Estimer un modèle de Cox en temps continu avec SPSS[1]

Gilbert Ritschard

Département d'économétrie, Université de Genève

http://mephisto.unige.ch

gilbert.ritschard@themes.unige.ch

octobre 2004

[1]Ce texte constitue un chapitre de la documentation sur « Modélisation des événements et transitions du parcours de vie (Event History Analysis): Mise en œuvre avec SPSS » proposée sur le site de l'institut PAVIE *http://www2.unil.ch/pavie/documentation/introduction.html*. Il a bénéficié d'une relecture attentive et de remarques constructives de Jean-Marie Le Goff et Yann Forney que nous remercions vivement. L'auteur assume cependant la responsabilité des erreurs qui pourraient subsister. Toutes les remarques sont les bienvenues.

Table des matières

1 Introduction **1**

2 Données illustratives **2**

3 Estimation d'un premier modèle simple **2**
3.1 Le dialogue Cox Regression 2
3.2 La syntaxe 4
3.3 L'output fourni 5
3.3.1 Interprétation des coefficients estimés 5
3.3.2 Autres résultats 5

4 Le traitement des attributs catégoriels **7**

5 Rappels théoriques **13**
5.1 Notations et rappel de quelques concepts 13
5.2 Cadre théorique du modèle de Cox 14
5.3 Deux mots sur la procédure d'estimation 16

6 Evaluation et sélection de modèles **18**
6.1 Significativité individuelle des coefficients 18
6.2 Evaluation globale 20
6.2.1 Le moins 2 log vraisemblance 21
6.2.2 Les statistiques du khi-deux 21
6.2.3 Statistique du rapport de vraisemblance 22
6.2.4 Pseudo R^2 23
6.3 Comparaison et sélection de modèles 24
6.3.1 Critères de comparaison 25
6.3.2 Estimation de modèles emboîtés avec SPSS 27
6.3.3 Sélection de modèles 29
6.4 Résidus partiels 33

7 Estimation de la fonction de survie **36**

8 Modèles stratifiés **38**

9 Modèle avec attributs variant dans le temps **41**

10 Test de l'hypothèse de proportionnalité **44**
10.1 Tests graphiques 44
10.2 Examen des résidus partiels 47

10.3 Tests statistiques 50
10.4 Quelques solutions en cas de rejet 52

11 Pour en savoir plus **53**

A Annexes **55**
A.1 Les données : contenu du fichier rrdat.1 55
A.2 Principe de l'estimation de $S_0(t)$ 59

Références **60**

Index **61**

1 Introduction

Le modèle de Cox (1972) en temps continu, connu aussi sous le nom de modèle semi-paramétrique à risques proportionnels, est un modèle d'analyse biographique (ou survie) du type régression. Il exprime le risque instantané, le *hazard rate*, de connaître l'événement étudié après une durée d'exposition donnée en fonction d'une combinaison linéaire de facteurs explicatifs. Il permet de quantifier et de tester les effets propres de caractéristiques individuelles telles que sexe, niveau de formation, classe sociale, nationalité, expérience passée, etc., sur le risque de transition.

Le modèle est conçu pour des données longitudinales individuelles qui se présentent sous forme d'événements datés permettant de caractériser des épisodes, les durées écoulées entre le début de l'exposition au risque et la réalisation de l'événement. Par exemple, si l'on s'intéresse au divorce, le début de l'exposition au risque est la date du mariage et la fin de l'épisode la date du divorce ou la fin de l'observation. Ainsi, si l'on dispose de panels de données comme ceux fournis par le panel suisse des ménages (http://www.swisspanel.ch), où le statut marital, par exemple, est représenté par une variable propre à chaque date, il conviendra de les transformer au préalable sous forme d'épisodes.

Ce guide se propose d'expliquer étape par étape à l'aide d'un exemple l'ajustement du modèle semi-paramétrique à risques proportionnels avec la procédure de régression de Cox de SPSS. On cherchera à démontrer toutes les possibilités offertes par la procédure, mais aussi ses limites. La présentation est organisée comme suit. La section 2 commence par présenter le jeu de données qui sert à l'illustration. La section 3 décrit la procédure, dialogue et syntaxe, pour estimer un modèle simple et propose un premier survol des résultats fournis. Le section 4 discute deux alternatives pour introduire des covariables catégorielles. La section 5 rappelle quelques notions de l'analyse des biographies et les fondements théoriques nécessaires pour apprécier la portée et les limites du modèle de Cox. Les indicateurs fournis par SPSS pour l'évaluation du modèle font l'objet de la section 6 où l'on discute également les procédures de sélection de modèles. A la section 7, on décrit comment obtenir les fonctions de survie prédites par le modèle. La section 8 est consacrée aux modèles stratifiés et la section 9 à la prise en compte de covariables variant dans le temps. Enfin, la section 10 présente plusieurs tests graphiques et statistiques de l'hypothèse de proportionnalité des risques sous-jacente au modèle de Cox. On termine à la section 11 avec des indications sur les sources, ouvrages et sites internet, où l'on trouvera des compléments utiles à ce guide.

Les données et les exemples de syntaxes présentés sont disponibles sur http://www2.unil.ch/pavie/documentation/introduction.html.

2 Données illustratives

Pour illustrer la présentation nous utilisons les données de Blossfeld et Rowher (2002). Deux raisons motivent ce choix. D'une part, le jeu de données est librement accessible sur le site de distribution du logiciel TDA (Rohwer et Pötter, 2002) http://steinhaus.stat.ruhr-uni-bochum.de/tda.html, ce qui permet à chacun de reproduire les résultats présentés ici. D'autre part, ce choix permet au lecteur de comparer directement les résultats fournis par SPSS avec ceux obtenus par les auteurs à l'aide de TDA.

Les données sont extraites de l'enquête biographique allemande réalisée entre 1981 et 1983 (Mayer et Brückner, 1989). Elles sont constituées de 600 épisodes relatifs à 201 individus choisis au hasard. Le phénomène que l'on se propose d'analyser est la durée d'emploi, ou plus précisément comment la fonction de risque de terminer un emploi est influencée par des caractéristiques individuelles.

Nous utilisons ici le fichier `rrdat.sav` au format SPSS que l'on trouve sur http://www2.unil.ch/pavie/documentation/introduction.htm. Celui-ci a été créé à partir du fichier texte `rrdat.1` mis à disposition sur le site TDA. Les données originales ainsi que la procédure pour les importer dans SPSS, y attacher des étiquettes et créer quelques variables supplémentaires est décrite dans l'annexe A.1. On admet dès maintenant que `rrdat.sav` est ouvert.

3 Estimation d'un premier modèle simple

Nous commençons par illustrer l'estimation d'un premier modèle dans lequel nous cherchons à expliquer le risque de terminer un emploi en fonction du nombre d'emplois précédents *pnoj*, du niveau d'éducation *edu* et du genre *sex*.

Comme toujours dans SPSS, on peut procéder soit par le menu et les dialogues soit par la syntaxe. Voyons d'abord le dialogue.

3.1 Le dialogue Cox Regression

Par le menu : Analysis/Survival/Cox Regression (Figure 1) on active le dialogue de la figure 2.

Par rapport à une régression classique ou logistique, on note que la réponse (variable dépendante) est ici caractérisée par deux variables : `Time`, la durée de l'épisode, soit `tfp` dans notre cas, et `Status`, la variable de troncature qui indique si l'épisode se termine avec l'événement considéré, la fin de l'emploi, ou s'il s'agit d'un épisode tronqué. Après avoir sélectionné la

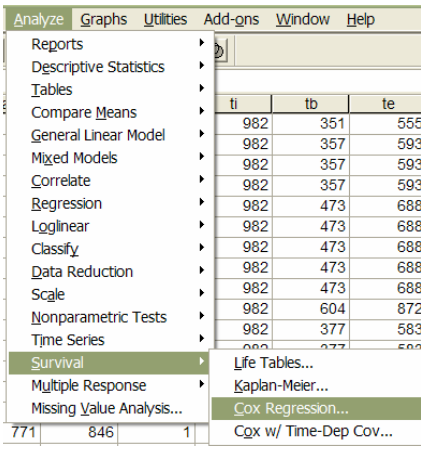

FIG. 1: Menu analyse de survie

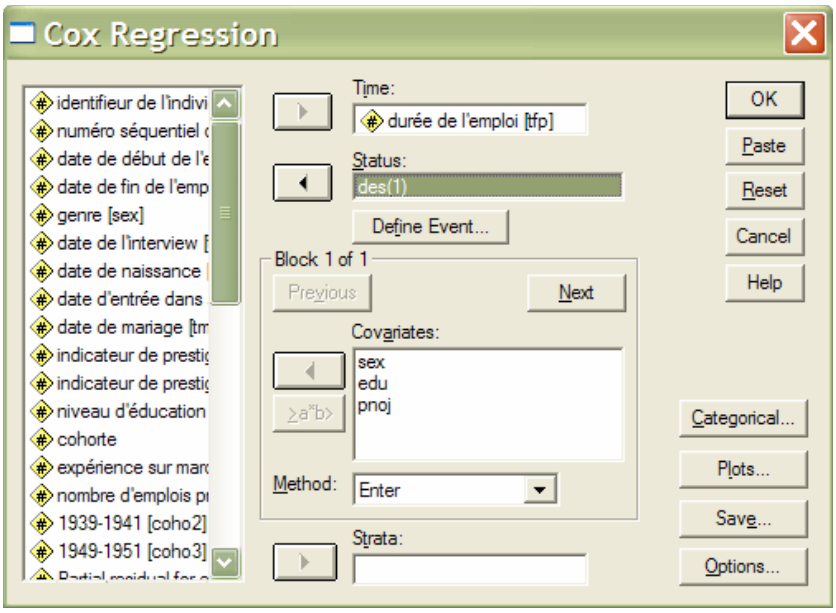

FIG. 2: Dialogue Régression de Cox

variable `Status`, *des* dans notre cas, on active le dialogue 3 en cliquant sur le bouton « Define Event » . Dans ce dialogue on indique la (ou les) valeur de la variable sélectionnée qui correspondent à une réalisation de l'événement, soit 1 ici puisque *des* est une variable binaire qui prend la valeur 1 en cas de fin d'emploi et 0 sinon (voir A.1 pour la construction de la variable).

Finalement, on sélectionne les variables explicatives dans le champs `Covariates` et on lance la procédure d'estimation en cliquant sur "OK".

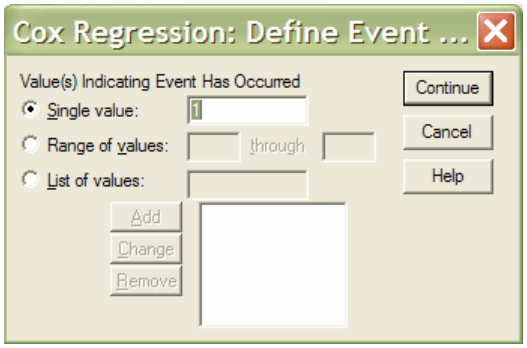

FIG. 3: Dialogue Définition de l'événement

3.2 La syntaxe

On peut également définir la même procédure sous forme de syntaxe. Le plus simple dans un premier temps est de générer celle-ci à partir du dialogue 2 en cliquant sur "Paste" plutôt que sur "OK". La syntaxe correspondant aux options choisies est alors générée dans une nouvelle fenêtre "Syntax" ou, s'il y en a déjà une d'ouverte, à la fin de celle-ci. Pour notre exemple on génère ainsi la commande

```
COXREG
   tfp  /STATUS=des(1)
   /METHOD=ENTER sex edu pnoj
   /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

`COXREG` est le nom de la commande SPSS. Il est suivi du nom de la variable durée qui définit la longueur des épisodes. La sous-commande obligatoire `STATUS` définit la variable de troncature et les covariables sont précisées avec `METHOD` qui est utilisé ici avec l'option `ENTER`. La sous-commande `CRITERIA` sert à définir des paramètres de contrôle que nous discuterons ultérieurement.

L'avantage de la syntaxe est que, contrairement au contenu des dialogues, elle peut être sauvegardée dans un fichier et réutilisée ultérieurement, ou encore être transmise à d'autres personnes. Par ailleurs, il est souvent plus simple, comme nous le ferons par la suite, d'éditer une syntaxe que de compléter à nouveau les dialogues nécessaires pour exécuter des variantes d'un premier modèle.

Pour exécuter la commande, il suffit de sélectionner la partie de la syntaxe que l'on veut exécuter et de cliquer sur . (Toute commande qui est partiellement sélectionnée sera exécutée !) Le menu Run permet de choisir entre exécuter tout le fichier, la partie sélectionnée, la commande courante ou depuis le curseur jusqu'à la fin.

3.3 L'output fourni

L'exécution de la commande, que cela soit par le dialogue ou la syntaxe, produit l'output reproduit à la figure 4.

3.3.1 Interprétation des coefficients estimés

L'estimation proprement dite du modèle est donnée dans le tableau "Variables in the Equation".

Sous B on y lit le coefficient estimé de chaque facteur explicatif. Celui-ci mesure l'effet du facteur sur le logarithme du risque. Par exemple, le coefficient de *pnoj* vaut $-.084$, ce qui signifie qu'à chaque nouvel emploi, le logarithme du risque diminue de $-.084$. Il est souvent plus aisé d'interpréter l'exponentiel du coefficient donné sous Exp(B) qui correspond à ce que d'autres logiciels nomment parfois *hazard ratio*. Pour le nombre d'emplois précédents, cet exponentiel vaut 0.919 et nous indique que, toutes choses égales par ailleurs, le risque est multiplié par 0.9 à chaque fois que *pnoj* augmente d'une unité. Pour le genre, l'exponentiel du coefficient est environ de 1.5. Ceci signifie que, pour un même niveau d'éducation et un même nombre d'emplois précédents, le risque de voir une femme ($sex = 2$) terminer son emploi est 1.5 fois celui d'un homme ($sex = 1$).

La colonne SE (Standard Error) donne l'erreur standard du coefficient qui mesure la variabilité de l'estimateur utilisé. Les trois colonnes 'Wald', 'df' et 'Sig' concernent la statistique de Wald utilisée pour tester la significativité individuelle de chaque coefficient. Elles donnent respectivement la valeur de la statistique, ses degrés de liberté et son degré de signification. On remarque ici que si les effets du genre (*sex*) et du nombre d'emplois précédents (*pnoj*) sont significatifs au seuil de 5%, le niveau d'éducation n'affecte lui pas significativement le risque de terminer un emploi et donc la durée de l'emploi.

3.3.2 Autres résultats

Le tableau "Case Processing Summary" nous indique que nos données sont constituées de 600 épisodes dont 142 (23.7%) sont tronqués (*Censored*). Aucune donnée n'a été supprimée. En particulier, on note qu'aucun épisode tronqué n'est plus court que la plus petite période observée entre début et fin d'emploi. Le cas échéant, ces épisodes qui n'entrent à aucun moment dans la population exposée au risque seraient ignorées par la procédure.

Les informations sous "Block 0 : Begining Block" concernent le modèle sans les facteurs explicatifs *sex, edu, pnoj*. En fait, on a ici simplement la

Cox Regression

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	458	76.3%
	Censored	142	23.7%
	Total	600	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		600	100.0%

a. Dependent Variable: tfp durée de l'emploi

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

-2 Log Likelihood
5169.140

Block 1: Method = Enter

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5142.326	26.834	3	.000	26.814	3	.000	26.814	3	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.402	.097	17.274	1	.000	1.496
edu	.018	.021	.765	1	.382	1.018
pnoj	-.084	.034	6.011	1	.014	.919

Covariate Means

	Mean
sex	1.420
edu	11.270
pnoj	1.487

FIG. 4: Output pour le 1er modèle simple

valeur du -2 log vraisemblance, -2 *Log Likelihood*, que l'on peut sommairement interpréter comme la distance entre les prédictions données par ce modèle 0 et les observations. Cette statistique, appelée aussi -2LogLik, ne nous dit rien en soi, mais est utile en tant que valeur de référence pour juger de l'apport des facteurs explicatifs introduits ultérieurement.

Les résultats du modèle spécifié figurent sous "Block 1". On y trouve le tableau déjà discuté des coefficients, mais aussi, sous "Omnibus Tests of Model Coefficients", des indicateurs globaux d'ajustement, soit le -2 log vraisemblance, la statistique du 'Score Test' ('Overall Score') et celle du rapport de vraisemblance donnée sous 'Change From Previous Step' et 'Change From Previous Block'. Ces deux dernières sont ici les mêmes, la différence entre étape (step) et bloc (block) apparaîtra ultérieurement. Nous discutons ces statistiques à la section 6. Retenons simplement ici qu'il s'agit de statistiques du khi-deux qui mesurent l'amélioration par rapport au modèle de référence, celui du bloc 0. Le gain est d'autant meilleur que ces statistiques sont grandes pour leurs degrés de liberté. En règle générale, on considère le gain comme significatif lorsque le degré de signification ('Sig') est inférieur à 5% (.05), ce qui est clairement le cas dans notre exemple. Nos trois facteurs permettent donc d'améliorer significativement l'ajustement par rapport au modèle sans facteurs explicatifs.

Enfin, l'output standard donne la valeur moyenne des variables explicatives. Il s'agit de la moyenne sur les 600 épisodes et non de la moyenne des 201 individus.

4 Le traitement des attributs catégoriels

Comme dans la régression classique, les variables explicatives catégorielles sont admises mais doivent être recodées sous forme d'un ensemble de variables auxiliaires.

Le recodage est essentiel. En effet, si l'on introduit une variable catégorielle sans la recoder, elle sera traitée comme quantitative ce qui implique en particulier que d'une part l'on tient compte de l'ordre de ses modalités, ce qui n'a par exemple aucun sens pour des variables purement nominales comme l'état-civil ou le domaine d'activité, et d'autre part, que les intervalles de valeurs sont comparables ce qui signifie, si l'on considère par exemple les catégories célibataire, marié, veuf, divorcé, que la différence entre marié et célibataire équivaut à celle entre divorcé et veuf, et que l'écart entre divorcé et célibataire vaut 3 fois celui entre célibataire et veuf. Ceci n'est évidemment pas défendable.

Dans certains cas, lorsque les catégories correspondent à des classes de

valeurs, des classes d'âges par exemple, traiter la variable comme quantitative peut se justifier. Cependant, il peut être avantageux de la traiter comme catégorielle pour capter notamment des non linéarités. C'est par exemple le cas pour la variable *cohorte* construite pour notre jeux de donnée (voir annexe A.1). En la traitant comme catégorielle, on autorise un rapport de risque différent entre les 3ème et 2ème cohortes qu'entre les 2ème et 1ère, ce qui n'est pas possible en la traitant comme quantitative.

De façon générale, une variable avec p catégories doit être recodée avec $p-1$ variables auxiliaires. Le codage le plus fréquent est sous forme de variables indicatrices prenant les valeurs 0 ou 1.

Prenons l'exemple de la variable *cohorte* qui compte trois catégories : 1="1929-1931", 2="1939-1941" et 3="1949-1951". On définit les variables indicatrices

$$\begin{array}{lll} coho1 = 1 & \text{si } cohorte = 1 & \text{et 0 sinon} \\ coho2 = 1 & \text{si } cohorte = 2 & \text{et 0 sinon} \\ coho3 = 1 & \text{si } cohorte = 3 & \text{et 0 sinon.} \end{array}$$

En fait ces trois variables sont linéairement dépendantes : il suffit de connaître la valeur de deux d'entre elles pour en déduire la valeur de la troisième. On a par exemple : $coho1 = 1 - coho2 - coho3$, ce qui en mots nous dit que si l'on est ni de la 2ème cohorte, ni de la 3ème, alors on appartient nécessairement à la 1ère. Deux variables auxiliaires suffisent donc pour coder les trois catégories de *cohorte*. Si l'on introduit les 3 variables dans la liste des facteurs explicatifs (ce que nous vous invitons à expérimenter), SPSS détecte la redondance et élimine arbitrairement la dernière des 3 variables.

Le recodage peut se faire soit en ajoutant les variables *coho2* et *coho3* à la base de données, soit directement dans le cadre de la procédure de régression de Cox.

Pour la définition individuelle de chaque variable indicatrice on peut utiliser le dialogue de recodage en une nouvelle variable auquel on accède par le menu "Transform/Recode/Into Different Variables". Le dialogue se présente comme illustré à la figure 5. On sélectionne d'abord le nom de la variable à recoder, *cohorte* dans notre cas, puis on précise le nom de la nouvelle variable, *coho2*, ainsi qu'éventuellement son étiquette, "1939-1941". Il faut ensuite cliquer sur 'Change' pour enregistrer le nom de la nouvelle variable. Pour le recodage proprement dit, on clique ensuite sur 'Old and New Values', ce qui active le dialogue représenté en dessous du premier. On choisit dans la partie gauche la (ou les) ancienne(s) valeur(s) que l'on veut coder 1, et sous 'New Values', on indique '1'. Il faut ensuite cliquer sur 'Add' pour ajouter cette définition à la liste avant de passer à la définition de la valeur suivante. On répète l'opération pour les autres valeurs cibles (0 dans notre cas, que l'on

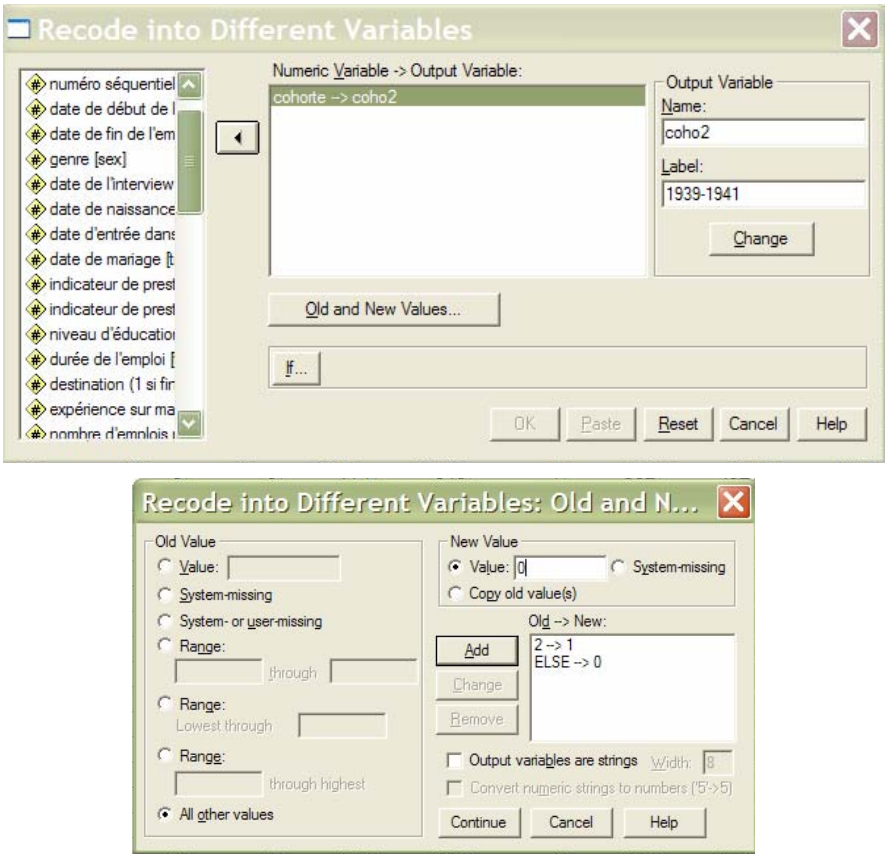

FIG. 5: Construction de la variable indicatrice *coho2*

associe à toutes les autres valeurs). Ne pas oublier de cliquer 'Add' avant de fermer le dialogue en sélectionnant 'Continue'. La variable *coho2* est finalement créée en cliquant sur 'OK' dans le premier dialogue. La même opération doit ensuite être répétée pour construire *coho3* et, le cas échéant, d'autres variables indicatrices.

Pour les utilisateurs un peu expérimentés, il peut être beaucoup plus simple de créer les variables indicatrices par la syntaxe avec, parmi d'autres, la commande `RECODE`. Les variables *coho2* et *coho3* et leurs étiquettes sont ainsi générées avec le code

```
RECODE cohorte (2=1) (else=0) INTO coho2
RECODE cohorte (3=1) (else=0) INTO coho3
VARIABLE LABEL
     coho2      '1939-1941'
    /coho3      '1949-1951'.
FORMAT coho2 coho3 (F3.0).
```

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5124.530	44.890	5	.000	44.610	5	.000	44.610	5	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.385	.097	15.813	1	.000	1.470
edu	.022	.022	1.009	1	.315	1.022
pnoj	-.078	.035	5.052	1	.025	.925
coho2	.402	.114	12.342	1	.000	1.494
coho3	.423	.119	12.717	1	.000	1.526

FIG. 6: Résultats avec les variables indicatrices *coho2* et *coho3*

Finalement, on peut mesurer l'effet de la cohorte en ajoutant les deux variables *coho2* et *coho3* à la liste des variables explicatives. Le nouvel output est donné à la figure 6. On observe que l'effet de la cohorte est représenté par deux coefficients. La catégorie non représentée sert de référence. Ainsi, le coefficient de *coho2* mesure l'effet, toute chose égale par ailleurs, d'appartenir à la 2ème cohorte plutôt qu'à la première. De même, le coefficient de *coho3* mesure l'effet d'appartenir à la 3ème cohorte plutôt qu'à la première. L'exponentiel du coefficient (tout à droite) mesure le ratio du risque de terminer l'emploi pour un individu de la cohorte concernée par rapport à un individu de même profil mais de la cohorte 1. Rappelons que, dans d'autres logiciels, cette information figure parfois sous le nom de *hazard ratio*.

SPSS offre la possibilité de recoder les variables catégorielles directement dans la procédure de régression de Cox. La méthode est la même que pour la régression logistique. Par le dialogue (figure 2 page 3), après avoir ajouter la variable catégorielle *cohorte* dans la liste des covariables (*Covariates*) on active le dialogue de recodage de la figure 7 en cliquant sur 'Categorical'. Dans ce dialogue, on sélectionne les variables catégorielles parmi les covariables introduites précédemment. Par défaut, le recodage par variables indicatrices (*Indicator*) est utilisé avec la dernière catégorie comme référence. Pour retrouver les résultats précédents, il faut définir la première modalité comme référence. Pour cela, il faut, après avoir sélectionné la variable concernée dans la liste de 'Categorical Covariates', choisir le bouton de radio 'First' et cliquer sur 'Change' pour enregistrer le changement. Notons au passage, que le dialogue permet de choisir d'autres recodages sous forme de contrastes. Nous ne discutons pas ces contrastes ici et renvoyons le lecteur à la documentation

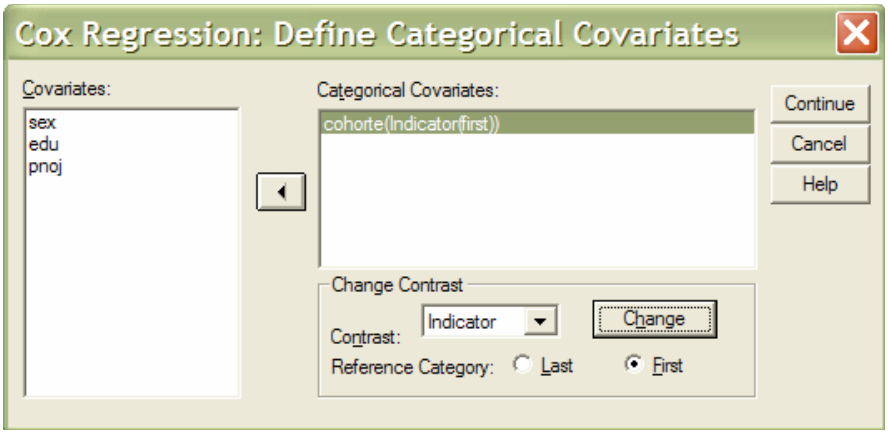

FIG. 7: Dialogue de recodage des variables catégorielles

SPSS Inc. (2003) pour leurs définitions. On ferme le dialogue en cliquant sur 'Continue', puis on lance la procédure d'estimation du modèle en cliquant 'OK' dans le premier dialogue. La syntaxe correspondante est

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator(1)
   /METHOD=ENTER sex edu pnoj cohorte.
```

Notons que la syntaxe plus souple permet de fixer n'importe quelle valeur de la variable comme catégorie de référence, contrairement au dialogue qui n'offre que le choix entre la première et la dernière.

La procédure ci-dessus donne les mêmes résultats que ceux obtenus avec les variables indicatrices *coho2* et *coho3* comme on peut le voir en comparant les outputs des figures 6 et 8.

L'avantage de la solution du recodage dans la procédure COXREG (option `CONTRAST`) est que l'on obtient ainsi une statistique de Wald pour la variable catégorielle en plus de celles pour les variables indicatrices. Sous l'hypothèse d'effet nul, la statistique de Wald est distribuée comme un khi-deux avec pour degrés de liberté le nombre de variables servant au codage. Pour la *cohorte*, on a Wald = 17.3 avec 2 degrés de liberté (un degré pour chaque variable auxiliaire nécessaire pour le recodage), ce qui est clairement significatif (Sig = .000).

L'inconvénient est qu'elle peut induire une certaine confusion quant à l'interprétation des variables auxiliaires servant au recodage et donc à l'interprétation des coefficients associés. Il convient pour cela de bien vérifier le codage dans le tableau "Categorical Variable Codings" (figure 9) avant de tenter toute interprétation. Dans cet exemple, on voit que la première

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5124.530	44.890	5	.000	44.610	5	.000	44.610	5	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.385	.097	15.813	1	.000	1.470
edu	.022	.022	1.009	1	.315	1.022
pnoj	-.078	.035	5.052	1	.025	.925
cohorte			17.312	2	.000	
cohorte(1)	.402	.114	12.342	1	.000	1.494
cohorte(2)	.423	.119	12.717	1	.000	1.526

FIG. 8: Résultats avec la variable catégorielle *cohorte*

Categorical Variable Codings[b]

		Frequency	(1)	(2)
cohorte[a]	1=1929-1931	226	0	0
	2=1939-1941	187	1	0
	3=1949-1951	187	0	1

a. Indicator Parameter Coding

b. Category variable: cohorte

FIG. 9: Codage utilisé de la variable *cohorte*

variable auxiliaire, colonne (1), est associée à la cohorte 2, et la deuxième, colonne (2), est associée à la cohorte 3.

La variable *sex* est catégorielle également, mais n'admet que deux valeurs successives 1 (homme) et 2 (femme). On peut évidemment aussi la définir comme catégorielle. Avec deux catégories cependant, une seule variable auxiliaire suffit au recodage qui ne revêt plus alors la même nécessité. Nous invitons le lecteur à vérifier qu'un codage indicateur avec 1 (homme) comme catégorie de référence donne le même résultat que sans recodage, tandis que si l'on prend comme référence 2 (femme), la dernière catégorie, le signe du coefficient change. Lorsque l'écart $v_2 - v_1$ entre les deux valeurs n'est pas unitaire, le recodage sous forme 0-1 amplifiera le coefficient associé en le multipliant par cet différence $v_2 - v_1$.

5 Rappels théoriques

5.1 Notations et rappel de quelques concepts

De façon générale, les épisodes sont caractérisés (Blossfeld et Rowher, 2002, p. 39) par des quadruplets (o_i, d_i, s_i, r_i), où o_i est l'état d'origine et d_i l'état de destination de l'épisode i, par exemple o_i = célibataire et d_i = marié, s_i et r_i étant respectivement la date de début et de fin d'épisode, soit par exemple la date de naissance (ou d'entrée en âge de se marier) et la date de mariage.

Les procédures de SPSS, et en particulier la procédure de régression de Cox, ne permettent de traiter que des événements simples, c'est-à-dire des cas où l'on n'a qu'un état d'origine possible (0), la destination étant soit l'événement a eu lieu (1) ou l'événement n'a pas eu lieu (0). Ces derniers cas, correspondent aux épisodes tronqués, soit ceux pour lesquels la période d'exposition au risque observée ne se termine pas par la réalisation de l'événement. De plus, SPSS ne prend pas en compte les dates de début et de fin d'épisode, mais seulement sa durée.

Pratiquement, dans SPSS les épisodes sont donc caractérisés, comme on l'a vu dans l'exemple introductif au point 3, par deux variables seulement, `Time` et `Status`

$$\begin{aligned} &\texttt{Time} \text{ (durée de l'épisode)} && : \quad t_i = r_i - s_i \\ &\texttt{Status} \text{ (variable de troncature)} && : \quad d_i = \begin{cases} 1 & \text{si l'événement a lieu} \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

Notons que pour la variable de troncature (*Status*), on peut indiquer (voir dialogue de la figure 3 page 4) une variable non nécessairement dichotomique en précisant les valeurs qui correspondent à la réalisation de l'événement. Par exemple, si l'on considère l'événement « entrer en union » et que l'on dispose d'une variable codée 1 pour union libre, 2 pour mariage et 0 sinon, on peut utiliser cette variable pour la troncature en indiquant que l'événement a lieu lorsqu'elle prend la valeur 1 ou 2.

Nous rappelons à présents quelques concepts de l'analyse de survie nécessaires à la compréhension du modèle semi-paramétrique à risque proportionnel de Cox.

Notons T la durée qui s'écoule depuis le temps 0 jusqu'à la réalisation de l'événement étudié (fin de l'emploi, mariage, naissance,...). A priori cette durée T, continue, est aléatoire et l'objectif est de caractériser sa distribution.

On peut pour cela utiliser indifféremment l'une des fonctions suivantes

$$\begin{array}{lcl}
\text{densité} & f(t) & = \lim_{dt\to 0} p(T \in [t, t+dt])/dt \\
\text{distribution} & F(t) & = p(T < t) \\
\text{survie} & S(t) & = p(T \geq t) = 1 - F(t) \\
\text{risque } (\textit{hazard}) & h(t) & = \lim_{dt\to 0} p(T \in [t, t+dt] \mid T \geq t)/dt = \dfrac{f(t)}{S(t)} \\
\text{risque } (\textit{hazard}) \text{ cumulé} & H(t) & = \displaystyle\int_{t_0}^{t} h(s)ds = \int_{t_0}^{t} \frac{f(s)}{S(s)} = -\ln S(t))
\end{array}$$

Dans le contexte de l'analyse de survie ou de l'analyse des biographies, il est d'usage de considérer soit la fonction de survie $S(t)$, soit la fonction de risque instantané $h(t)$, soit encore parfois la fonction de risque cumulé $H(t)$. Par exemple, l'estimateur de Kaplan-Meier donne une estimation non paramétrique de $S(t)$ et celui de Nelson-Aalen de $H(t)$. Précisons que le risque instantané $h(t)$ est une densité conditionnelle. C'est l'équivalent continu de la probabilité de connaître l'événement en t, sachant qu'on ne l'a pas connu avant et que l'on a donc «survécu» jusqu'en t.

5.2 Cadre théorique du modèle de Cox

Les modèles de type régression, dont le modèle de Cox étudié ici fait partie, postulent que la durée dépend d'une forme linéaire

$$\mathbf{x}'\beta = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p \tag{1}$$

de p caractéristiques individuelles $x_1, x_2, \ldots, x_p$. Ils comprennent deux grandes familles : les modèles à occurrences accélérées (*accelarated failure times*) qui expriment directement la durée en fonction de la forme linéaire des caractéristiques, $T(\mathbf{x}) = T_0 \exp(\mathbf{x}'\beta)$, et les modèles à risques proportionnels

$$\begin{array}{rcl}
h(t, \mathbf{x}) & = & h_0(t) \exp(\mathbf{x}'\beta) \\
\ln h(t, \mathbf{x}) & = & \ln h_0(t) + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p
\end{array} \tag{2}$$

où c'est le risque instantané qui dépend de la forme linéaire $\mathbf{x}'\beta$. Le modèle de Cox relève de cette seconde catégorie.

Lorsqu'on interprète les résultats d'un modèle, il est évidemment primordial de savoir quel type de modèle a été estimé. Dans notre exemple du point 3, la variable *pnoj*, nombre d'emplois précédents, influence négativement la variable réponse qui est dans ce cas le risque. Comme la durée et le risque évoluent en sens contraire (plus le risque de terminer l'emploi est

faible, plus sa durée est longue), l'effet de *pnoj* sur la durée est évidemment de sens contraire.[1] Il s'agit évidemment de ne pas confondre.

Les modèles de la forme (2) sont dits à risques proportionnels car, avec cette écriture, pour deux individus avec profils respectifs $\mathbf{x}$ et $\mathbf{z}$, le rapport des risques (*hazard ratio*)

$$hr(\mathbf{x};\mathbf{z}) = \frac{h(t,\mathbf{x})}{h(t,\mathbf{z})} \quad = \quad \exp\big((\mathbf{x}-\mathbf{z})'\beta\big) \tag{3}$$

est indépendant de la durée t.

Notons que si le rapport des risques reste ainsi constant, le risque lui-même ne l'est pas nécessairement. L'évolution du risque avec le temps est ici reflétée par le seul terme $h_0(t)$ qui représente la fonction de risque de référence (*baseline hazard*).

Le modèle de Cox de base est un modèle à risques proportionnels. Pour l'output de la figure 4 et en posant $\mathbf{z} = (0,0,0)$, le modèle estimé peut ainsi, en retenant la forme (3), s'écrire

$$\widehat{hr}\big(sex, edu, pnoj; 0\big) \quad = \quad e^{.402\,sex}\, e^{.018\,edu}\, e^{-.084\,pnoj}$$

avec $hr\big(sex, edu, pnoj; 0\big) = h(t, sex, edu, pnoj)/h_0(t)$ où $h_0(t) = h(t,0,0,0)$.

Il s'agit d'un modèle semi-paramétrique. La partie paramétrique concerne les effets des caractéristiques sur le risque qui sont paramétrés par la forme linéaire (1). L'aspect non-paramétrique a trait à l'absence d'hypothèses sur la forme de la fonction de risque $h_0(t)$, ou de façon équivalente, sur la distribution de la durée T_0 pour le profil de référence.

Le fait que le modèle ne nécessite pas d'hypothèse sur la forme de la distribution des durées a largement contribué à sa popularité. L'inconvénient est qu'on ne dispose pas alors d'expression analytique simple pour prédire un risque ou une durée pour un profil donné. Des prédictions sont néanmoins possibles par le biais de techniques non paramétriques similaires au principe de Kaplan-Meier (voir section 7). Notons simplement ici que pour un modèle à risque proportionnel de la forme (2), la fonction de survie s'écrit (voir par exemple Hosmer et Lemeshow, 1999, pp 92-93)

$$S(t,\mathbf{x}) \quad = \quad S_0(t)^{\exp(\mathbf{x}'\beta)} \quad . \tag{4}$$

Le modèle s'appuie cependant sur l'hypothèse de proportionnalité des risques. Il convient donc de s'assurer de la pertinence de cette hypothèse. Nous montrons comment procéder à la section 10, où nous discutons également des possibilités d'assouplir cette contrainte de proportionnalité.

[1] L'intensité de l'effet est en règle générale aussi un peu différente.

5.3 Deux mots sur la procédure d'estimation

Le modèle est estimé en maximisant une fonction de vraisemblance partielle. Soit la suite ordonnée des durées différentes qui se sont écoulées jusqu'à la réalisation de l'événement

$$t_{(1)} < t_{(2)} < \cdots < t_{(q)} \ .$$

La vraisemblance partielle considère pour chaque cas non tronqué la probabilité qu'en $t_{(i)}$, ce soit précisément l'épisode i qui se termine plutôt qu'un des autres cas restant exposés en $t_{(i)}$. Sous l'hypothèse d'absence d'égalités entre durées (toutes les durées sont différentes), cette probabilité est

$$LP_i \ = \ \frac{h_{(i)}}{\sum_{j\geq i} h_{(j)}} \ = \ \frac{\exp(\mathbf{x}'_{(i)}\beta)}{\sum_{j\geq i} \exp(\mathbf{x}'_{(j)}\beta)} \ .$$

La vraisemblance partielle que l'on maximise est donc :

$$\begin{aligned} LP \ &= \ \prod_{i\in I_{\text{non tronc}}} \frac{\exp(\mathbf{x}'_{(i)}\beta)}{\sum_{j\geq i} \exp(\mathbf{x}'_{(j)}\beta)} \\ \ln LP \ &= \ \sum_{i\in I_{\text{non tronc}}} \left(\mathbf{x}'_{(i)}\beta - \ln\big(\sum_{j\geq i} \exp(\mathbf{x}'_{(j)}\beta)\big) \right) \ . \end{aligned} \tag{5}$$

Bien qu'il ne s'agisse pas d'une véritable vraisemblance, cette vraisemblance partielle se comporte et s'utilise comme une vraisemblance. Les estimations des coefficients β fournies par SPSS sont celles qui maximisent l'expression ci-dessus. Il n'y a pas d'expression analytique des estimateurs, si bien que le calcul des solutions s'appuie sur des procédures numériques dont on peut contrôler les critères d'arrêts avec l'option `CRITERIA`. Par exemple, en ajoutant la ligne

```
/CRITERIA BCON(1E-4) LCON(1E-5) ITERATE(20)
```

la procédure itérative s'arrête lorsque, entre deux itérations, la variation de tous les coefficients est inférieure à `BCON`= 10^{-4}, ou quand la variation de la vraisemblance partielle est inférieure à `LCON`= 10^{-5}, ou encore quand le nombre d'itérations atteint 20. Les valeurs données ici pour ces trois critères sont les valeurs par défaut.

On note que la vraisemblance partielle est indépendante de la fonction de référence $h_0(t)$ et donc de ses paramètres. Ceux-ci ne peuvent donc pas être estimés par cette approche.

Les erreurs standard des coefficients se déduisent de l'inverse de la matrice d'information $\mathbf{I}(\beta) = -\partial^2 LP/\partial\beta^2(\beta)$ évaluée en $\beta = \hat{\beta}$. L'inverse $\mathbf{I}(\hat{\beta})^{-1}$ est

en effet l'estimation de la matrice des variances et covariances des coefficients $\hat{\mathbf{V}}(\hat{\beta})$.

La matrice d'information est également à la base de deux statistiques de test dont l'usage sera discuté à la section 6. Il s'agit des statistiques du 'Score Test' et de Wald

$$\text{Score} \;=\; \mathbf{u}(0)'\,\mathbf{I}(0)^{-1}\,\mathbf{u}(0) \tag{6}$$

$$\text{Wald} \;=\; \hat{\beta}'\,\hat{\mathbf{V}}(\hat{\beta})^{-1}\,\hat{\beta} \;=\; \hat{\beta}'\,\mathbf{I}(\hat{\beta})\,\hat{\beta} \tag{7}$$

où $\mathbf{u}(\beta) = \partial LP/\partial\beta(\beta)$ désigne le vecteur des scores. Sous l'hypothèse H_0 : $\beta = 0$, ces deux statistiques sont distribuées selon un khi-deux avec le nombre de coefficients β comme degrés de liberté. La statistique de Wald s'adapte aisément au test de significativité d'un sous-ensemble de paramètres. Il suffit de ne considérer pour cela que le sous-vecteur de paramètres concernés et la partie correspondante de la matrice $\mathbf{I}(\hat{\beta})$. Par exemple, pour un seul paramètre β_j, la statistique de Wald s'écrit

$$\text{Wald}(\beta_j) \;=\; \frac{\beta_j^2}{\widehat{\text{Var}}(\hat{\beta}_j)} \;=\; \left(\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}\right)^2 \tag{8}$$

ce qui n'est rien d'autre que le carré du ratio critique coefficient sur erreur standard, souvent appelé t de Student.

Les données réelles ne vérifient que rarement l'hypothèse de durées toutes différentes. En particulier lorsque les durées sont mesurées sous forme discrète, comme dans notre exemple où les durées d'emploi sont mesurées en mois, le nombre d'épisodes se terminant après une même durée peut être assez conséquent. Le calcul exact de la vraisemblance devient difficile dans ce cas, et l'on utilise alors des approximations. SPSS utilise l'approximation de Breslow qui consiste, lorsque plusieurs épisodes se terminent en $t_{(i)}$, à définir la contribution de chacun en utilisant le même dénominateur. Formellement, en notant $\mathcal{D}_i$ l'ensemble des indices des épisodes se terminant en $t_{(i)}$, la contribution de chaque épisode k de $\mathcal{D}_i$ est

$$\frac{h_{(k)}}{\sum_{j\geq i} h_{(j)}} \quad \text{pour tout } k \in \mathcal{D}_i \;\; .$$

Parmi les différentes propositions d'approximation faites dans la littérature, celle de Breslow est la plus simple. Ce n'est cependant pas la plus appropriée. Elle a tendance en effet à sous-estimer la valeur absolue des coefficients, le biais restant toutefois faible tant que le nombre d'événements se produisant simultanément reste petit. Pour plus de détails sur ce point, voir Therneau et

Grambsch (2000, pp 48-53). Contrairement à d'autres logiciels, SPSS n'offre pas d'autres choix. En présence de nombreuses égalités, il peut être préférable de recourir à une modélisation discrète, comme le modèle de Cox en temps discret (modèle logistique à cotes proportionnelles).

6 Evaluation et sélection de modèles

La pertinence statistique d'un modèle se fonde en règle général sur la significativité statistique individuelle des coefficients, sur l'ajustement global, et sur l'analyse des résidus. Pour le modèle de Cox, SPSS fournit divers indicateurs pour les deux premiers aspects. En ce qui concerne les résidus, SPSS ne permet d'obtenir que les résidus partiels de Schoenfeld associés à chaque covariable. Nous discutons brièvement ci-après ces divers indicateurs.

6.1 Significativité individuelle des coefficients

Comme on l'a vu par exemple à la figure 4 page 6, la procédure de régression de Cox donne pour chaque coefficient la statistique de Wald associée, ses degrés de liberté et son degré de significativité.

La procédure ne donne pas par contre le ratio critique, souvent appelé t de Student, entre le coefficient et son erreur standard que l'on examine traditionnellement pour juger de sa significativité statistique. Une règle sommaire avec le ratio critique est de considérer le coefficient comme significatif lorsque le ratio est supérieur à 2.

Pour un coefficient individuel, la statistique de Wald est simplement le carré du ratio critique (eq. 8 page 17). Par exemple, dans l'output de la figure 8 page 12, le coefficient de *sex* est 0.385 et son erreur standard 0.097. Le ratio critique vaut $0.385/0.097 = 3.969$ dont le carré est $3.696^2 = 15.8$, soit (aux erreurs d'arrondis près) la valeur indiquée pour la statistique de Wald. Selon la règle sommaire du ratio critique supérieur à 2, cette valeur de Wald est ici grande puisque supérieure $2^2 = 4$. Elle indique donc un effet significatif du genre (*sex*).

Pour une règle plus précise, on compare la valeur de Wald pour un coefficient à un chi-deux à 1 degré de liberté. En effet, sous l'hypothèse d'effet nul, la statistique de Wald d'un coefficient est distribuée comme un khi-deux à 1 degré de liberté. La significativité de la variable *sex* est ainsi confirmée par son degré de signification quasi nul

$$\text{Sig}(sex) = p(\chi^2_{(1dl)} > 15.8) \simeq .000 \ .$$

L'avantage de la statistique de Wald par rapport au ratio critique est qu'elle se généralise au test simultané de la significativité de plusieurs coefficients, ce qui permet d'attribuer une valeur de la statistique aux variables catégorielles dont l'effet est mesuré par plus d'un coefficient. Sous l'hypothèse d'effet nul, la statistique de Wald est dans ce cas distribuée comme un khi-deux dont le nombre de degrés de liberté est égal au nombre de variables auxiliaires utilisées pour la représenter. Ainsi, pour la variable *cohorte* à la figure 8 page 12, la statistique de Wald vaut 17.3. Comme la variable est codée avec deux variables indicatrices, on a 2 degrés de liberté. Le degré de signification, c'est-à-dire ici la probabilité que sous l'hypothèse d'effet nul on obtienne une valeur de Wald supérieure à la valeur obtenue de 17.3, vaut

$$\text{Sig}(cohorte) = p(\chi^2_{(2dl)} > 17.3) \simeq .000 \ .$$

Ceci nous dit qu'il y a une probabilité quasi nulle que l'effet mesuré soit dû au seul hasard. L'effet de la variable est donc significatif.

A côté de cette valeur globale pour la variable *cohorte*, on a naturellement aussi une statistique de Wald propre à chaque variable auxiliaire servant à son codage. Ici (figure 8), on a 12.3 pour *cohorte*(1) qui est associé à la cohorte 2, et 12.7 pour *cohorte*(2) associé à la cohorte 3. Chacune de ses valeurs est clairement grande (> 4) et les deux coefficients sont donc significativement non nuls. En fait, c'est la différence entre l'effet respectivement des cohorte 2 et 3 par rapport à la cohorte 1 qui est significative. Un changement de codage donne lieu évidemment à d'autres coefficients individuels. La statistique de Wald associé à la variable catégorielle reste cependant inchangé. On illustre ce dernier point en prenant comme référence la dernière cohorte plutôt que la première. Par rapport à la syntaxe de la page 11, il suffit, pour sélectionner la dernière catégorie comme référence, de supprimer le '(1)' après `Indicator` dans la sous-commande `CONTRAST` (ou de le remplacer par '(3)')

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator
   /METHOD=ENTER sex edu pnoj cohorte.
```

Les seuls changements par rapport aux résultats précédents (figures 8 et 9) apparaissent dans les tableaux reportés en figure 10. On remarque, que cette fois seul un des deux coefficients mesurant l'impact de *cohorte* est significatif. La variable auxiliaire *cohorte*(1) est associée ici à la première cohorte et *cohorte*(2) à la deuxième. Les effets mesurés sont les écarts par rapport à la cohorte 3. Ainsi, le premier coefficient (écart de cohorte 1 par rapport à 3) est simplement l'opposé du troisième coefficient obtenu précédemment (écart de cohorte 3 par rapport à 1). Le deuxième par contre mesure l'écart entre

Categorical Variable Codings[b]

		Frequency	(1)	(2)
cohorte[a]	1=1929-1931	226	1	0
	2=1939-1941	187	0	1
	3=1949-1951	187	0	0

a. Indicator Parameter Coding

b. Category variable: cohorte

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.385	.097	15.813	1	.000	1.470
edu	.022	.022	1.009	1	.315	1.022
pnoj	-.078	.035	5.052	1	.025	.925
cohorte			17.312	2	.000	
cohorte(1)	-.423	.119	12.717	1	.000	.655
cohorte(2)	-.021	.122	.030	1	.864	.979

FIG. 10: Codage indicateur de *cohorte* avec cohorte 3 comme référence

la cohorte 2 et la cohorte 3. Ce coefficient n'est pas significatif, ce qui signifie que, toutes choses égales par ailleurs, il n'y pas de différence significative entre les individus de la cohorte 1939-41 et ceux de la cohorte 1949-51. On note que, comme annoncé, la statistique de Wald pour l'effet cohorte ne change pas et reste donc très significative. Cette significativité est due ici à la forte différence entre la première cohorte et les deux autres.

6.2 Evaluation globale

Les indications sur l'ajustement global du modèle sont données dans le tableau intitulé "Omnibus Tests of Models Coefficients". On y trouve :

- le moins 2 log vraisemblance ('-2 Log Likelihood') qui correspond ici à -2 fois le logarithme du maximum atteint de la vraisemblance partielle,
- la statistique du khi-deux du Score Test ('Overall'),
- la statistique du khi-deux du rapport de vraisemblance ('Change from Previous Step').

Le moins 2 log vraisemblance mesure la qualité de reproduction des données ou si l'on veut la 'distance' entre les prédictions fournies par le modèle et les observations. Les deux autres mesures évaluent le gain par rapport au modèle naïf m_0 qui ne tient compte d'aucun facteur explicatif ($\beta = 0$). C'est en quelque sorte une 'distance' entre le modèle estimé et le modèle naïf.

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5124.530	44.890	5	.000	44.610	5	.000	44.610	5	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 1. Method = Enter

FIG. 11: Indicateurs d'ajustement global pour modèle de la figure 10

6.2.1 Le moins 2 log vraisemblance

Le moins 2 log vraisemblance, appelé ci-après -2LogLik et noté $-2LL$, mesure la qualité de reproduction des données ou si l'on veut la 'distance' entre les prédictions fournies par le modèle et les observations. Pour notre exemple, il vaut $-2LL = 5124.53$.

L'ajustement du modèle est d'autant meilleure que sa vraisemblance est forte, et que donc son -2LogLik est petit. Ici on observe par exemple que l'ajustement du modèle, $-2LL(m) = 5124.53$ est meilleur que celle du modèle sans facteurs explicatifs, $-2LL(m_0) = 5169.14$. Cette dernière valeur correspond au -2 log vraisemblance initial indiqué sous le tableau à la figure 11. On trouve aussi cette valeur sous Block 0 dans l'output de la figure 4. En termes absolus, on ne sait pas quand un -2LogLik est petit ou quand il est grand. Cette quantité n'a de sens que lorsqu'on la compare avec le -2LogLik d'autres modèles. En particulier, nous verrons que le khi-deux du rapport de vraisemblance compare la valeur du -2LogLik observé de 5124.53 au 5169.14 du -2LogLik du modèle du Block 0.

6.2.2 Les statistiques du khi-deux

Ces statistiques permettent d'évaluer si globalement l'ensemble des facteurs explicatifs considérés améliore significativement l'ajustement du modèle naïf qui ne tient compte d'aucun facteur. En d'autres termes, pour un modèle avec p coefficients, ils permettent de tester l'hypothèse $H_0 : \beta_1 = 0, \ldots, \beta_p = 0$ contre l'hypothèse qu'un au moins des coefficients est non nul. Par exemple, pour le modèle de la figure 10 on teste simultanément les $p = 5$ coefficients. Trois statistiques du khi-deux sont utilisables dans ce contexte, soit

- Score Test (Score)
- Wald (Wald)
- rapport de vraisemblance (G^2).

Sous l'hypothèse H_0 (modèle naïf correct), les trois statistiques sont distribuées asymptotiquement selon une loi du khi-deux à p degrés de liberté. On

considère donc l'amélioration par rapport au modèle naïf comme significative lorsque la valeur de ces statistiques est suffisamment grande, soit lorsque leur degré de signification est inférieur, en règle générale, à 5%. Le degré de signification est ici défini comme la probabilité que le khi-deux prenne une valeur supérieure à la valeur observée de la statistique. A la figure 11, la valeur du Score est 44.89. Ceci est grand pour $p = 5$ degrés de liberté,[2] ce que confirme le degré de signification

$$\text{Sig}(Score) = p(\chi^2_5 > \text{Score}_{obs} = 44.89) \simeq .000 \ .$$

Le modèle naïf doit donc être rejeté au profit du modèle ajusté. Cela ne signifie pas que le modèle ajusté est satisfaisant, mais nous dit simplement que le modèle ajusté fait mieux que le modèle naïf.

Les autres statistiques donnent des valeurs du même ordre de grandeur que le Score. SPSS ne donne pas la valeur de la statistique de Wald pour le test global, mais fournit la valeur du rapport de vraisemblance G^2 sous 'Change from Previous Step'. On a pour notre exemple $G^2 = 44.61$, ce qui est effectivement très proche de la valeur de 44.89 du Score.

Le test du rapport de vraisemblance présente l'avantage de pouvoir être très facilement exploité pour la comparaison de modèles emboîtés dès que l'on dispose des -2LogLik. Nous explicitons ci-dessous sa logique.

6.2.3 Statistique du rapport de vraisemblance

Soit m_2 un modèle ajusté et m_1 un modèle emboîté dans m_2, dit parfois aussi modèle contraint. Le modèle m_1 est contraint s'il s'obtient en imposant des contraintes sur les paramètres de m_2. Pour simplifier, pensons simplement à des contraintes de nullité du type $\beta_j = 0$. Comme m_2 est plus complet (plus de paramètres libres), sa vraisemblance est plus grande que celle de m_1 qui n'est qu'un cas particulier de m_2. Dès lors, le rapport de vraisemblance partielle $LP(m_2)/LP(m_1)$ ne peut être que plus grand ou égal à 1. L'idée du test du rapport de vraisemblance est de voir si le rapport est significativement plus grand que 1, auquel cas le modèle réduit m_1 sera considéré comme significativement différent de m_2.

On a l'équivalence

$$\frac{LP(m_2)}{LP(m_1)} > 1 \quad \Leftrightarrow \quad -2LL(m_1) + 2LL(m_2) > 0 \tag{9}$$

[2] Rappelons que l'espérance du khi-deux est son nombre de degrés de liberté, $\text{E}(\chi^2_p) = p$, et que sa variance vaut deux fois ce nombre $\text{Var}(\chi^2_p) = 2p$.

et l'on regarde alors de façon équivalente si la quantité $-2LL(m_1)+2LL(m_2)$ est significativement non nulle. C'est cette quantité

$$G^2(m_2|m_1) = -2LL(m_1) + 2LL(m_2) \tag{10}$$

que l'on appelle *statistique du rapport de vraisemblance*. Elle est aussi connue sous le nom de déviance. Sous l'hypothèse que la spécification m_1 est correcte, et sous certaines conditions de régularité, elle est asymptotiquement distribuée comme un khi-deux avec pour degrés de liberté le nombre de contraintes qui permet de passer de la spécification m_2 au modèle m_1.

Dans le cas de l'évaluation globale, les seuls modèles considérés sont le modèle ajusté m et le modèle naïf m_0. On obtient m_0 en imposant la nullité de tous les p coefficients, soit $\beta_j = 0, j = 1, \ldots, p$. La statistique $G^2(m_0|m)$ qui compare m à m_0 a donc p degrés de liberté, comme nous l'avons déjà dit plus haut. C'est la valeur de cette statistique que SPSS donne sous 'Change from Previous Block'. Dans le cas particulier, il s'agit du changement entre les blocs 0 et 1. On vérifie en effet que l'on a

$$G^2(m_0|m) = -2LL(m_0) - 2LL(m) = 5169.14 - 5124.53 = 44.61 \ .$$

Comme on a $p = 5$, la significativité de ce gain par rapport au modèle naïf est

$$\mathrm{Sig}\big(G^2(m_0|m)\big) = p(\chi^2_5 > 44.61) \simeq .000 \ .$$

La probabilité étant négligeable, le gain est clairement significatif.

6.2.4 Pseudo R^2

Bien qu'ils ne soient pas calculés par SPSS, les pseudo R^2 se calculent aisément à partir des -2LogLik. Ils s'interprètent plus ou moins comme la proportion de réduction du défaut d'ajustement ou "dispersion résiduelle" du modèle naïf, la dispersion étant mesurée en terme de -2LogLik.

Les pseudo R^2 les plus pertinents sont dans notre contexte ceux de Cox et Snell (1989) et de Nagelkerke (1991)

$$\text{Cox-Snell } R^2_{CS} = 1 - \exp\Big(-\frac{-2LL(m_0) + 2LL(m)}{n}\Big) \tag{11}$$

$$\text{Nagelkerke } R^2_N = \frac{R^2_{CS}}{1 - \exp\big(2LL(m_0)/n\big)} \ . \tag{12}$$

Le n est le nombre de termes dans la vraisemblance, soit en général le nombre d'observations. Pour le modèle de Cox, il s'agit du nombre d'événements, puisque seuls les épisodes se terminant avec la réalisation de l'événement

donnent lieu à un terme dans la vraisemblance partielle. On trouve ce nombre, soit $n = 458$, dans le premier tableau de la page 6. Dans la formule (11), le numérateur de l'argument de l'exponentiel est, au signe près, la statistique du rapport de vraisemblance. Pour le modèle m de la figure 10 on trouve, comme on l'a vu, cette valeur dans la partie d'output 11. On obtient ainsi :

$$R^2_{CS} \quad = 1 - \exp\left(\frac{-44.61}{458}\right) = 1 - e^{-.0974} = 1 - .907 = .093 \ .$$

Le R^2_N de Nagelkerke s'en déduit, en divisant cette valeur par

$$1 - \exp\left(\frac{1}{n} 2LL(m_0)\right) = 1 - \exp\left(\frac{-5169.14}{458}\right) = 1 - e^{-11.3} \simeq 1 \ .$$

Dans cet exemple, on a donc $R^2_N \simeq R^2_{CS} = 9.3\%$. La part "expliquée" de la "dispersion" totale est de l'ordre de 9%.

6.3 Comparaison et sélection de modèles

Lors de la modélisation, on est en général confronté à plusieurs variantes de spécifications, c'est-à-dire, de choix de variables explicatives. Par exemple, le modèle avec les variables *sex*, *edu* et *pnoj* est-il satisfaisant, ou convient-il de tenir compte également de l'effet *cohorte* et peut-être d'autres variables encore ? Du point de vue purement statistique, qui ne saurait toutefois être le seul à être pris en compte, il s'agit de voir si l'ajout d'un bloc de variables, dans notre cas l'ajout de la variable catégorielle *cohorte*, améliore significativement la qualité d'ajustement du modèle. On doit donc comparer des modèles.

La comparaison peut se faire soit manuellement sur la base des informations fournies par l'estimation individuelle de chaque modèle, soit en exploitant les possibilités d'estimation de modèles emboîtés. Le choix du modèle peut également s'appuyer sur une procédure de sélection pas à pas des facteurs explicatifs pertinents.

Dans SPSS, ces diverses approches sont liées au choix de la méthode ('Method') et de la définition des blocs (dialogue 2 page 3). Dans la syntaxe, ces choix se font par la sous-commande `METHOD` de `COXREG`. On peut ainsi spécicier :

- un modèle individuel
- des modèles emboîtés (Blocks)
- une procédure de sélection pas à pas (Steps).

Nous discutons ces diverses options du point de vue de la comparaison de modèles.

6.3.1 Critères de comparaison

Nous commençons par décrire les critères de comparaison que l'on peut établir à partir des informations que l'on obtient en estimant individuellement les modèles.

La spécification d'un modèle individuel se fait en choisissant la méthode `ENTER`, soit dans le dialogue (figure 2), ou comme illustré par exemple dans la syntaxe page 11. C'est le choix par défaut.

Rapport de vraisemblance entre deux modèles Lorsque les modèles sont emboîtés, on peut les comparer avec la statistique du rapport de vraisemblance. Selon la formule (10) page 23, les valeurs de la statistique se calculent aisément à partir des -2LogLik ou LogLik que pratiquement tous les logiciels, et en particulier SPSS, donnent dans leurs outputs. Il existe aussi des formes des statistiques du Score et de Wald pour ce type de comparaison, mais elles nécessitent des informations que l'on ne trouve pas en général dans les outputs fournis.

Considérons à titre d'exemple les modèles

m_1 avec facteurs explicatifs *sex, edu, pnoj*
m_2 avec facteurs explicatifs *sex, edu, pnoj, cohorte*

où *cohorte* est catégorielle. Pour m_1, le -2LogLik est donné dans l'output 4 page 6. Il vaut $-2LL(m_1) = 5142.326$. Celui de m_2 est donné dans l'output 11 page 21. Il vaut $-2LL(m_2) = 5124.53$. On en déduit par différence la valeur de la statistique du rapport de vraisemblance comparant m_1 à m_2

$$G^2(m_1|m_2) \;=\; 5142.326 - 5124.53 \;=\; 17.796 \;. \tag{13}$$

La valeur de la statistique peut également s'obtenir par différence des statistiques du rapport de vraisemblance comparant respectivement m_2 et m_1 au modèle naïf m_0, soit pour notre exemple

$$\begin{aligned} G^2(m_1|m_2) &= G^2(m_0|m_2) - G^2(m_0|m_1) \\ &= 44.61 - 26.814 \;=\; 17.796 \;. \end{aligned}$$

Le modèle m_2 a deux coefficients de plus que m_1. Ce nombre définit les degrés de liberté associés à la statistique. Notons que c'est aussi la différence des degrés de liberté de $G^2(m_0|m_2)$ et $G^2(m_0|m_1)$, soit $5 - 3 = 2$.

La valeur de 17.796 pour 2 degrés de liberté est grande. Elle est par exemple largement supérieure au seuil critique pour un risque de première espèce de 5% qui vaut $\chi^2_2(.95) = 5.99$. Les modèles sont donc clairement différents, d'où il ressort que la variable *cohorte* améliore significativement

la qualité d'ajustement. On peut relever que la valeur de la statistique est très proche de celle de la statistique de Wald associée à la variable *cohorte* qui vaut 17.312 (figure 10 page 20). Ceci s'explique par le fait que les deux modèles comparés ne diffèrent que par cette variable *cohorte*. Le rapport de vraisemblance n'est donc ici qu'une façon alternative de tester l'effet de *cohorte*. Les deux statistiques étant distribuées selon la même loi du χ^2, il n'est pas étonnant qu'elles prennent des valeurs proches.

De façon générale, le test du rapport de vraisemblance est réputé être plus puissant que celui de Wald, et devrait être préféré dans des cas litigieux. Par ailleurs, précisons que la procédure décrite ci-dessus n'est évidemment pas limitée à la comparaison de modèles qui ne diffèrent que par une variable. Elle est par contre limitée à la comparaison de modèles emboîtés.

Les critères AIC et BIC. Comme le souligne notamment Raftery (1995), le degré de signification (*p-value*) en général, et donc en particulier de la différence entre deux modèles emboîtés, perd de son intérêt quand le nombre d'observations devient grand, le moindre écart devenant statistiquement significatif. Une alternative permettant de pallier ce défaut, consiste à exploiter les critères d'information AIC et BIC. Ces critères se prêtent également à la comparaison de modèles non emboîtés. Ils ont donc une portée plus générale que le test du rapport de vraisemblance, et semblent mieux indiqués pour la comparaison de modèles.

Bien que SPSS ne calcule pas les AIC et BIC, on peut très facilement les obtenir à partir des informations que l'on trouve dans les outputs.

Le critère d'information AIC, dû à Akaike (1983), et le critère bayésien BIC, introduit par Schwarz (1978) et popularisé par Raftery (1995), tiennent compte à la fois de la qualité de l'ajustement mesurée par le -2LogLik et de la complexité du modèle mesurée par le nombre p de paramètres. Dans notre cas p est simplement le nombre de coefficients. Les critères AIC et BIC se définissent simplement en ajoutant une pénalisation pour la complexité au -2LogLik

$$\text{AIC} = -2LL(m) + 2p \tag{14}$$
$$\text{BIC} = -2LL(m) + \log(n)\, p \ . \tag{15}$$

La pénalisation pour la complexité est indépendante du nombre n d'observations (épisodes se terminant par la réalisation de l'événement) pour le AIC, tandis qu'elle croît avec $\log(n)$ pour le BIC. La pénalisation est plus forte pour le BIC que pour le AIC.

La valeur des ces critères n'a pas de signification *per se*. Elle n'a de sens que par comparaison avec la valeur d'autres modèles. Pratiquement, on re-

tient le modèle qui prend la plus petite valeur du critère, soit celui qui réalise le meilleur compromis entre parcimonie et qualité ajustement.

Pour le modèle $m2$, par exemple, le calcul se fait comme suit

$$\begin{aligned} \text{AIC} &= 5124.53 + \underbrace{2 \cdot 5}_{10} = 5134.53 \\ \text{BIC} &= 5124.53 + \log(458) \cdot 5 = 5124.53 + \underbrace{6.13 \cdot 5}_{30.65} = 5155.18 \ . \end{aligned}$$

Le tableau 1 récapitule les valeurs des AIC et BIC des modèles m_0 (naïf), m_1 et m_2. Parmi ces trois modèles, le dernier m_2 est le plus satisfaisant puisqu'il donne lieu aussi bien au plus petit AIC qu'au plus petit BIC.

TAB. 1: Pseudo R^2, AIC et BIC

	-2LogLik	LR	p	R^2_N	AIC	BIC
m_0	5169.140	0	0	0	5169.1	5169.1
m_1	5142.326	26.814	3	0.057	5148.3	5160.7
m_2	5124.530	44.610	5	0.093	5134.5	5155.2

Il est naturel cependant de se demander à partir de quel écart, un modèle peut être considéré comme clairement supérieur à un autre. Raftery (1995) préconise la règle suivante pour un écart $\text{BIC}(m_1) - \text{BIC}(m_2)$ positif.

inférieur à 2	non significatif
entre 2 et 6	légère supériorité de m_2
entre 6 et 10	supériorité marquée de m_2
plus de 10	supériorité évidente de m_2.

Dans notre cas, on a $\text{BIC}(m_1)$ - $\text{BIC}(m_2) = 4.5$ et le modèle m_2 qui inclut la variable *cohorte* n'est donc que légèrement supérieur au modèle sans *cohorte*. Ceci nuance la conclusion du test du rapport de vraisemblance pour lequel la différence est clairement significative.

On pourrait imaginer une règle similaire pour le AIC. Il nous paraît plus important, toutefois, de relever que ce dernier est connu pour pénaliser insuffisamment la complexité. Nous conseillons donc de privilégier le BIC.

6.3.2 Estimation de modèles emboîtés avec SPSS

Pour la comparaison de modèles emboîtés, SPSS offre la possibilité de calculer directement les statistiques du rapport de vraisemblance nécessaires, leurs degrés de liberté et leurs signification en spécifiant dans une même commande les modèles emboîtés que l'on veut comparer.

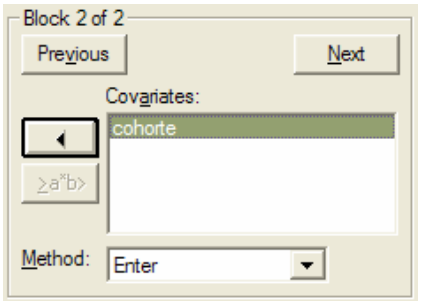

FIG. 12: Ajout d'un bloc à la spécification de la figure 2

La spécification de modèles emboîtés se fait par le dialogue (figure 2 page 3) en définissant plusieurs blocs. On définit un nouveau bloc en cliquant sur 'Next' dans le cadre 'Block i of k' (figure 12). Par la syntaxe, les blocs sont définis en incluant plusieurs sous-commandes `METHOD` avec option `ENTER` dans une même commande `COXREG`

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator
   /METHOD=ENTER sex edu pnoj
   /METHOD=ENTER cohorte.
```

Les blocs sont utilisés de façon incrémentale. SPSS estime tout d'abord le modèle correspondant au 1er bloc, puis, sous Block 2, le modèle constitué de la réunion des variables spécifiées sous bloc 1 et bloc 2. De même, sous Block 3, on a le modèle constitué de la réunion des variables des blocs 1 à 3, et ainsi de suite. Par exemple, en ajoutant un second bloc avec la variable *cohorte* et en définissant celle-ci comme catégorielle avec la dernière cohorte comme référence (cf syntaxe ci-dessus), on obtient au bloc 1 le même résultat qu'à la figure 4 page 6, et au bloc 2 l'output illustré à la figure 13.

Le tableau 'Variables in the Equation' est le même que celui obtenu avec le modèle incluant directement toutes les variables (figure 10 page 20). Le tableau 'Omnibus Tests of Model Coefficients' diffère lui de celui de la figure 11 qui correspond au même modèle. La différence porte sur la statistique du rapport de vraisemblance (sous 'Change from Previous Block'). En effet, la valeur reportée ici, soit 17.795, compare le modèle du bloc 2 à celui du bloc précédent 1, c'est-à-dire le modèle m_2 avec *sex, edu, pnoj* et *cohorte* à celui, m_1, avec les seules variables *sex, edu* et *pnoj*, tandis que dans le cas précédent on compare le modèle complet au modèle naïf. La valeur donnée est, à un arrondi près, la valeur calculée précédemment en (14).

L'avantage par rapport au calcul manuel est que SPSS nous donne le

Block 2: Method = Enter

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5124.530	44.890	5	.000	17.795	2	.000	17.795	2	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 2. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.385	.097	15.813	1	.000	1.470
edu	.022	.022	1.009	1	.315	1.022
pnoj	-.078	.035	5.052	1	.025	.925
cohorte			17.312	2	.000	
cohorte(1)	-.423	.119	12.717	1	.000	.655
cohorte(2)	-.021	.122	.030	1	.864	.979

FIG. 13: Output supplémentaire résultant de l'ajout du bloc 2

degré de signification, ce qui évite d'avoir à consulter une table du χ^2. Pour notre exemple, on a $Sig = p(\chi^2_2 > 17.795) = .000$, ce qui est négligeable et confirme que l'ajout de *cohorte* améliore significativement le modèle m_1.

6.3.3 Sélection de modèles

Un des objectif de la comparaison de modèles, est de sélectionner le plus pertinent statistiquement. Dans cette optique SPSS propose des heuristiques de sélection pas à pas de modèles. On sélectionne la méthode ('Method') dans le dialogue principal (figure 14) ou par la syntaxe.

SPSS propose des méthodes de sélection avant ('Forward') et arrière ('Backward').

La méthode 'Forward' commence avec le modèle sans variables explicatives et cherche, parmi la liste des covariables indiquées, la plus significative. Si le degré de signification de la variable est inférieur à une valeur d'entrée spécifiée, la variable est retenue et la procédure continue. Sinon la procédure s'arrête. A l'étape suivante, on cherche à nouveau la variable la plus significative parmi celles qui ne sont pas déjà sélectionnées. Après l'ajout d'une nouvelle variable, la procédure vérifie que les variables précédemment introduites ne deviennent pas non significatives selon le seuil de sortie indiqué. Le cas échéant les variables devenues non significatives sont exclues du modèle. La procédure est ensuite répétée jusqu'à ce qu'on ne puisse plus compléter le modèle avec un nouvel effet significatif.

La méthode 'Backward' procède en sens inverse. Elle commence avec toutes les variables et élimine successivement les moins significatives. Seules évidemment sont éliminées des variables non significatives selon le seuil de sortie indiqué. Lorsqu'il n'y a plus de variables non significatives, la procédure s'arrête.

SPSS propose trois variantes 'Forward' et trois variantes 'Backward'. Les trois variantes 'Forward' sélectionnent les variables à introduire sur la base de la statistique du 'Score'. Elles se différencient par la statistique de test utilisée pour éliminer le cas échéant les variables non significatives. Les options sont 'Conditional', 'LR' et 'Wald'. La première, 'Conditional', utilise une statistique du rapport de vraisemblance fondée sur des estimations conditionnelles des paramètres, la seconde, 'LR', utilise un rapport de vraisemblance partielle et la troisième la statistique de 'Wald'. Les mêmes options sont disponibles avec la procédure 'Backward'.

Le choix de la méthode se fait dans le dialogue ou dans la syntaxe en

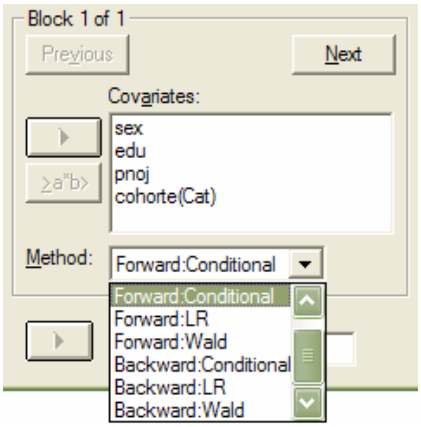

FIG. 14: Choix d'une procédure de sélection pas à pas

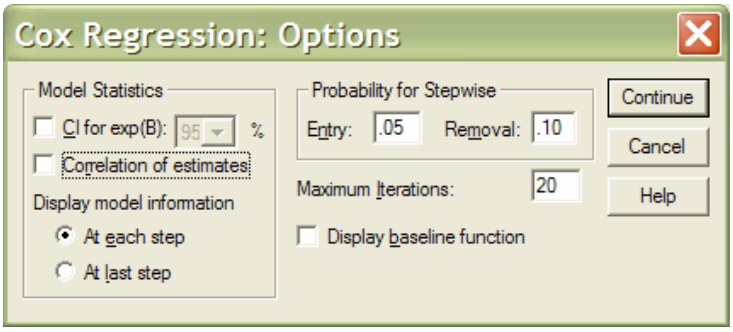

FIG. 15: Les options

Omnibus Tests of Model Coefficients[d,e]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	5149.491	20.240	1	.000	19.649	1	.000	19.649	1	.000
2[b]	5130.851	38.706	3	.000	18.640	2	.000	38.289	3	.000
3[c]	5125.507	43.772	4	.000	5.345	1	.021	43.634	4	.000

a. Variable(s) Entered at Step Number 1: sex

b. Variable(s) Entered at Step Number 2: cohorte

c. Variable(s) Entered at Step Number 3: pnoj

d. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

e. Beginning Block Number 1. Method = Forward Stepwise (Conditional LR)

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	sex	.423	.095	19.956	1	.000	1.527
Step 2	sex	.408	.095	18.479	1	.000	1.503
	cohorte			18.190	2	.000	
	cohorte(1)	-.458	.118	15.161	1	.000	.632
	cohorte(2)	-.079	.120	.433	1	.511	.924
Step 3	sex	.372	.096	15.042	1	.000	1.451
	pnoj	-.078	.035	5.085	1	.024	.925
	cohorte			17.099	2	.000	
	cohorte(1)	-.429	.118	13.149	1	.000	.651
	cohorte(2)	-.040	.121	.111	1	.739	.961

Variables not in the Equation[a,b,c]

		Score	df	Sig.
Step 1	edu	.838	1	.360
	pnoj	6.088	1	.014
	cohorte	18.439	2	.000
	cohorte(1)	17.943	1	.000
	cohorte(2)	3.917	1	.048
Step 2	edu	1.049	1	.306
	pnoj	5.096	1	.024
Step 3	edu	1.010	1	.315

a. Residual Chi Square = 24.421 with 4 df Sig. = .000

b. Residual Chi Square = 6.095 with 2 df Sig. = .047

c. Residual Chi Square = 1.010 with 1 df Sig. = .315

Model if Term Removed

Term Removed		Loss Chi-square	df	Sig.
Step 1	sex	19.649	1	.000
Step 2	sex	18.208	1	.000
	cohorte	18.640	2	.000
Step 3	sex	14.863	1	.000
	pnoj	5.346	1	.021
	cohorte	17.562	2	.000

FIG. 16: Output de la sélection 'Forward' des variables

Omnibus Tests of Model Coefficients[c,d]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	5124.530	44.890	5	.000	44.610	5	.000	44.610	5	.000
2[b]	5125.507	43.772	4	.000	.976	1	.323	43.634	4	.000

a. Variable(s) Entered at Step Number 1: sex edu pnoj cohorte

b. Variable Removed at Step Number 2: edu

c. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

d. Beginning Block Number 1. Method = Backward Stepwise (Likelihood Ratio)

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	sex	.385	.097	15.813	1	.000	1.470
	edu	.022	.022	1.009	1	.315	1.022
	pnoj	-.078	.035	5.052	1	.025	.925
	cohorte			17.312	2	.000	
	cohorte(1)	-.423	.119	12.717	1	.000	.655
	cohorte(2)	-.021	.122	.030	1	.864	.979
Step 2	sex	.372	.096	15.042	1	.000	1.451
	pnoj	-.078	.035	5.085	1	.024	.925
	cohorte			17.099	2	.000	
	cohorte(1)	-.429	.118	13.149	1	.000	.651
	cohorte(2)	-.040	.121	.111	1	.739	.961

Model if Term Removed

Term Removed		Loss Chi-square	df	Sig.
Step 1	sex	15.625	1	.000
	edu	.976	1	.323
	pnoj	5.308	1	.021
	cohorte	17.795	2	.000
Step 2	sex	14.862	1	.000
	pnoj	5.345	1	.021
	cohorte	17.561	2	.000

Variables not in the Equation[a]

		Score	df	Sig.
Step 2	edu	1.010	1	.315

a. Residual Chi Square = 1.010 with 1 df Sig. = .315

FIG. 17: Output de la sélection ‘Backward’ des variables

indiquant `FSTEP(...)` ou `BSTEP(...)` sur la sous-commande `METHOD`. Les options à préciser dans la parenthèse sont `COND`, `LR` et `Wald`. Les seuils pour les probabilités d’entrée et de sortie sont à préciser dans le sous-dialogue ‘Options’ (figure 15) respectivement sous ‘Entry’ et ‘Removal’ dans le cadre ‘Probability for Stepwise’. Avec la syntaxe, on les précise avec les options PIN (entrée) et POUT (sortie) de la sous-commande `CRITERIA` comme illustré ci-dessous

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator
   /METHOD=FSTEP(COND) sex edu pnoj cohorte
   /CRITERIA=PIN(.05) POUT(.10) .
```

Pour une variable catégorielle définie comme telle dans la procédure, les variables indicatrices qui la codent sont introduites ou supprimées en bloc. C'est le cas de la variable *cohorte* dans notre exemple.

La figure 16 montre l'output obtenu avec la syntaxe ci-dessus qui demande une sélection 'Forward', et la figure 17 celui que produit la sélection 'Backward' (en remplaçant `FSTEP` par `BSTEP`).

6.4 Résidus partiels

SPSS permet d'obtenir les résidus partiels de Schoenfeld du modèle de Cox estimé. Les résidus sont calculés pour les cas non tronqués seulement. Pour chaque cas, il y a p résidus, soit un résidu pour chaque coefficient β_j.

Contrairement aux résidus que l'on rencontre habituellement en régression, ces résidus partiels ne sont pas des écarts entre valeurs observées et prédites de la variable réponse. Les résidus portent sur les covariables et mesurent des écarts par rapport au profil moyen en termes de ces variables. Plus précisément, ils mesurent la différence entre la valeur des covariables d'un cas connaissant l'événement (la fin de l'emploi) en t et une moyenne pondérée des valeurs de ces mêmes covariables pour les cas exposés en t.

Comment les obtenir ? On les obtient en sélectionnant 'Partial Residuals' dans le dialogue 'Save' que l'on active depuis le dialogue principal de la régression de Cox (figure 18). Par la syntaxe, on obtient ces résidus avec la sous-commande `SAVE` de `COXREG`

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator
   /METHOD=ENTER sex edu pnoj cohorte
   /SAVE=PRESID.
```

Les résidus sont introduits comme variables supplémentaires dans la base de données active ('Working Space'). Par défaut, elles sont nommées `PRk_m`, où k est le numéro du coefficient, par exemple 3 pour celui associé à *pnoj*, et m sert à repérer le modèle. On peut les consulter avec l'éditeur de données (figure 19) et les analyser avec les outils statistiques et graphiques de SPSS. La racine *PR* du nom peut être changée avec la syntaxe. Par exemple

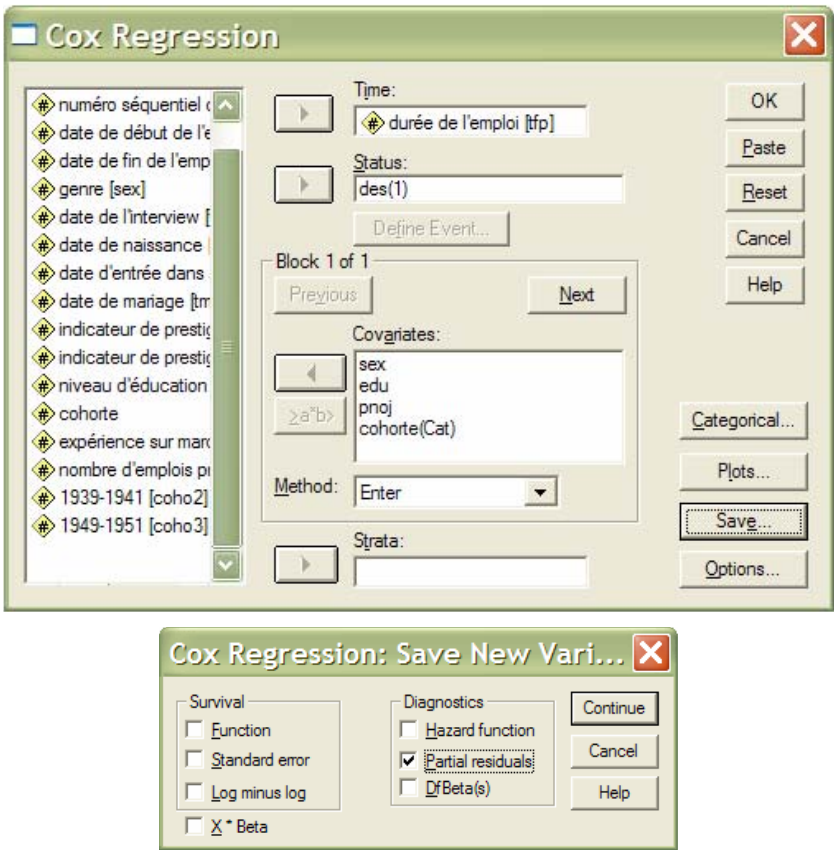

FIG. 18: Demande des résidus de Schoenfeld par le dialogue

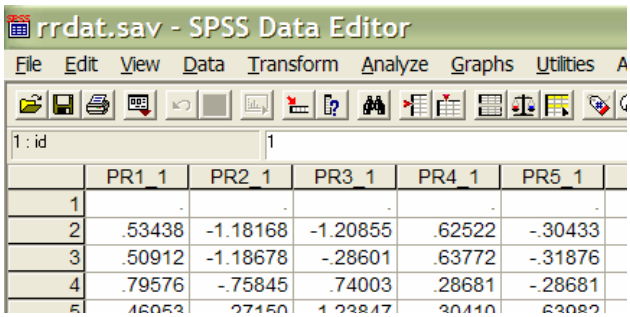

FIG. 19: Résidus partiels de Schoenfeld

`/SAVE=PRESID(sch)` génère les variables *sch*1, *sch*2, ... Attention, il convient de s'assurer que ces noms de variables n'existent pas déjà, sinon les résidus ne seront pas générés.

A quoi correspondent-ils ? Dans une régression classique, tout comme en régression logistique, les résidus ont trait en général à la différence entre la valeur observée de la variable dépendante et sa valeur prédite, cette dernière étant une estimation de sa valeur moyenne pour les cas avec un même profil. Dans notre contexte de modèles d'analyse de biographies, la réponse n'est pas un nombre, mais la fonction de risque instantané $h(t, \mathbf{x})$. La notion même de résidu en tant que différence entre une valeur observée et une valeur prédite n'est donc plus applicable. Le modèle de Cox étant un modèle à risques proportionnels, on pourrait songer, en retenant la spécification (3) page 15, à comparer le rapport de risque prédit au rapport effectif pour chaque cas. Ceci n'est cependant pas possible non plus, car le rapport effectif n'est pas observable. La définition de résidus est donc quelque chose de compliqué dans ce contexte d'analyse de biographies.

Plusieurs types de résidus ont été proposés dans la littérature : déviance, martingale, du score et de Schoenfeld (cf par exemple Therneau et Grambsch, 2000, chap. 4). Les résidus partiels de Schoenfeld sont les seuls que calcule SPSS, et donc les seuls que nous discutons ici.

Pour chaque cas non tronqué i, caractérisé par le temps t_i écoulé jusqu'à l'événement, la fin de l'emploi par exemple, et son profil $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$, soit le genre, le niveau d'éducation, le nombre d'emplois précédents et la cohorte d'appartenance, on définit autant de résidus qu'il y a de coefficients β, soit 5 dans notre exemple.

Chaque résidu r_{ij} est défini comme l'écart

$$r_{ij} \quad = \quad x_{ij} - \bar{x}_j(t_i, \hat{\beta}) \tag{16}$$

entre la valeur prise par i pour la covariable j et la moyenne pondérée $\bar{x}_j(t_i, \hat{\beta})$ des valeurs de cette covariable j pour l'ensemble des cas exposés au risque en t_i. Plus précisément on a $\bar{x}_j(t_i, \hat{\beta}) = \sum_{k \in \mathcal{R}_i} w_k(\hat{\beta}) x_{kj}$ où $\mathcal{R}_i$ est l'ensemble des cas exposés en t_i. Les pondérations $w_i(\hat{\beta})$ sont proportionnelles aux rapports de risques estimés $\widehat{hr}(\mathbf{x}_i, 0)$, soit $w_i(\hat{\beta}) = \exp(\mathbf{x}_i'\hat{\beta}) / \sum_{k \in \mathcal{R}_i} \exp(\mathbf{x}_k'\hat{\beta})$. Avec cette définition, la somme des résidus est nulle par construction pour chaque covariable j.

Un résidu partiel de Schoenfeld porte donc non pas sur la variable réponse, mais sur une covariable, soit *sex*, *edu*, *pnoj*, *cohorte*(1) ou *cohorte*(2) dans notre modèle. Elle indique dans quelle mesure le cas considéré qui a subi l'événement en t_i, s'écarte de la valeur moyenne de la covariable de tous les cas encore exposés en t_i.

Comme nous le verrons à la section 10, l'examen de ces résidus s'avère en particulier utile pour juger de la pertinence de l'hypothèse de proportionnalité. Contrairement aux résidus usuels que l'on rencontre en régression, ils ne

donnent pas d'indication sur la qualité de l'ajustement.

7 Estimation de la fonction de survie

Pour un modèle à risque proportionnel, la fonction de survie $S(t, \mathbf{x})$ d'un individu avec profil $\mathbf{x}$ se déduit de la fonction de survie de $S_0(t)$, correspondant au profil virtuel de référence $\mathbf{x} = 0$, par la relation

$$S(t, \mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\beta)} \ .$$

Le modèle de Cox donne l'estimation $\hat{\beta}$ des coefficients β. Pour obtenir une estimation de la probabilité de survie $S(t, \mathbf{x})$, il nous faut encore une estimation de la fonction $S_0(t)$ de référence.

En l'absence d'hypothèses sur la forme de la distribution, on estime $S_0(t)$ de façon non paramétrique, avec un estimateur très similaire à celui de Kaplan-Meier. Le lecteur intéressé trouvera une description du principe de cette estimation dans l'annexe A.2.

Fonction de survie pour profil donné. Avec la procédure `COXREG` de SPSS, on peut :

1. Générer le graphique de la fonction de survie $S(t, \mathbf{x})$ pour un profil donné $\mathbf{x}$. On peut aussi obtenir la fonction $H(t, \mathbf{x})$ de risque cumulé ('Cumulated Hazard').
2. Obtenir pour chaque épisode i non tronqué, l'estimation de la probabilité de survivre $S(t_i)$ au moment t_i où l'événement s'est produit.

Les graphiques s'obtiennent en sélectionnant les cases correspondantes dans le dialogue 'Cox Regression: Plots' que l'on active en cliquant sur 'Plots' dans le dialogue principal. Voir figure 20. Alternativement, on peut les obtenir par la syntaxe en ajoutant la sous-commande `PLOT` suivante à la commande `COXREG` :

```
/PLOT SURVIVAL HAZARD
```

Par défaut on obtient la courbe demandée pour le profil moyen $\bar{\mathbf{x}}$. Pour obtenir la courbe pour un autre profil, il faut préciser les valeurs des covariables dans le dialogue comme illustré à la figure 20. L'exemple montre les valeurs à fixer pour obtenir la fonction de survie de référence $S_0(t)$. On note que pour la variable catégorielle, nous avons donné la valeur utilisée comme catégorie de référence. Par la syntaxe, cette opération se fait en ajoutant la sous-commande `PATTERN` suivante

```
/PATTERN sex(0) edu(0) pnoj(0) cohorte(3)
```

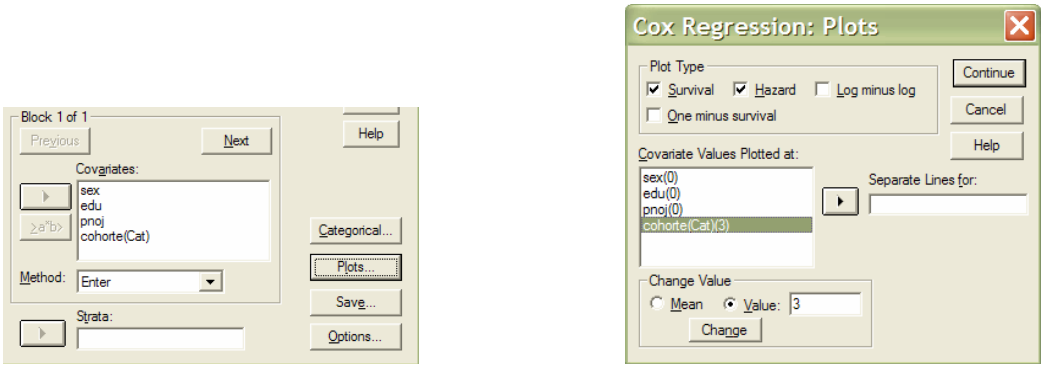

FIG. 20: Demande du plot des courbes de survie et de risque cumulé

Les probabilités de survie estimées en fin d'épisode s'obtiennent par le biais du dialogue 'Save', ou, par la syntaxe avec la sous commande `SAVE`. Par exemple, en complétant la commande `COXREG` avec la ligne

```
/SAVE SURVIVAL XBETA
```

on génère les probabilités $S(t_i, \mathbf{x}_i)$ de survie (avec `SURVIVAL`) et les valeurs de $(\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\beta}$ (avec `XBETA`). Elles sont ajoutées à la base de données active, par défaut sous les noms `SUR_m` et `XBE_m`, où m est un compteur qui prend la valeur de 1 au premier essai. On peut définir d'autres noms en les indiquant entre parenthèses, par exemple `SURVIVAL(surv)` et/ou `XBETA(rbeta)`.

Ces informations peuvent être utilisées pour divers calculs. A titre d'illustration, la syntaxe ci-dessous montre comment générer les courbes de survie $S(t, \bar{\mathbf{x}})$ et de risque cumulé $H(t, \bar{\mathbf{x}})$ pour le profil moyen $\bar{\mathbf{x}}$ à partir des deux

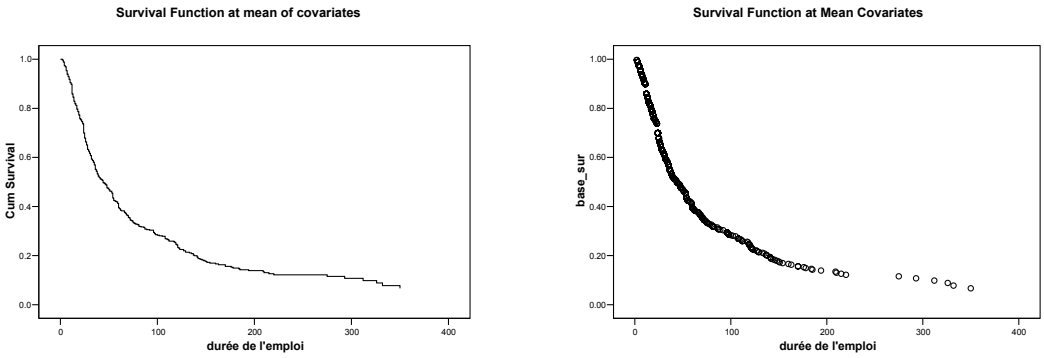

FIG. 21: Courbes générées par `COXREG` à gauche et par `GRAPH` à droite

variables sauvegardées précédemment.[3] Les calculs réalisés sont

$$\begin{aligned} S(t_i, \bar{\mathbf{x}}) &= S(t_i, \mathbf{x})^{1/\exp\left((\mathbf{x}-\bar{\mathbf{x}})'\hat{\beta}\right)} \\ H(t_i, \bar{\mathbf{x}}) &= -\ln S(t_i, \bar{\mathbf{x}}) \end{aligned}$$

```
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=indicator
   /METHOD=ENTER sex edu pnoj cohorte
   /PATTERN sex(0) edu(0) pnoj(0) cohorte(3)
   /PLOT SURVIVAL HAZARD
   /SAVE=SURVIVAL XBETA.

** Calcul et plot des fonctions de survie S et
** de risque cumulé H pour le profil moyen xbar.

COMPUTE base_sur = sur_1**(1/EXP(xbe_1)). /* S(t,x=xbar) */
COMPUTE base_haz = -LN(base_sur).         /* H(t,x=xbar) */

FILTER BY des.    /* Plot pour cas non tronqués seulement */
GRAPH
  /SCATTERPLOT(BIVAR)=tfp WITH base_sur
  /TITLE='Survival Function at Mean Covariates'.
GRAPH
  /SCATTERPLOT(BIVAR)=tfp WITH base_haz
  /TITLE='Cumulative Hazard Function at Mean Covariates'.
FILTER OFF.
```

La figure 21 montre les courbes de survie $S(t_i, \bar{\mathbf{x}})$ obtenues, à gauche, avec la sous-commande `PLOT` de `COXREG`, et à droite avec `GRAPH`.

8 Modèles stratifiés

`COXREG` permet d'estimer des modèles de Cox stratifiés. Un modèle stratifié est un modèle où le risque de référence (*baseline hazard*) est propre à chaque strate. Il ne comporte cependant qu'un jeu de coefficients β qui sont donc supposés être les mêmes pour toutes les strates.

[3] Pour un profil $\mathbf{x}^*$ autre que le profil moyen, on a $S(t_i, \mathbf{x}^*) = S(t_i, \mathbf{x})^{1/\exp\left((\mathbf{x}-\mathbf{x}*)'\hat{\beta}\right)}$. Comme $(\mathbf{x} - \mathbf{x}^*)'\hat{\beta} = (\mathbf{x} - \bar{\mathbf{x}})'\hat{\beta} - (\mathbf{x}^* - \bar{\mathbf{x}})'\hat{\beta}$, où le premier terme s'obtient avec `/SAVE = XBETA` et le second se calcule avec la commande `MATRIX - END MATRIX` en récupérant les coefficients $\hat{\beta}$ préalablement sauvegardés dans un fichier SPSS avec la sous-commande `/OUTFILE = COEFF(filename)`.

La stratification se fait selon la variable que l'on spécifie dans le champs 'Strata' du dialogue principal, ou par la syntaxe avec la sous-commande `STRATA`. Cette variable est supposée catégorielle. On repère ses valeurs par l'indice $s = 1, 2, \ldots$. Par exemple, une stratification selon le genre donnera lieu à 2 strates et une stratification selon la cohorte à 3 strates.

Lorsqu'on a précisé une variable de stratification, le modèle de Cox est donc estimé en admettant un risque de référence $h_{s,0}(t)$ différent pour chaque strate s. L'hypothèse générale de proportionnalité est alors remplacée par les deux hypothèses suivantes

1. Risques proportionnels pour individus d'une même strate.
2. Effets (β) des covariables indépendants de la strate.

La première de ces hypothèses, qui autorise des risques non proportionnels entre individus de strates différentes, est un relâchement de l'hypothèse générale de proportionnalité. La seconde postule que les effets des covariables sur le rapport de risque intra-strate sont les mêmes dans toutes les strates, ce qui permet d'estimer un seul jeu de coeffcients pour l'ensemble des strates.

Sous ces hypothèses, la vraisemblance partielle que l'on maximise devient

$$\begin{aligned} LP_{str} &= \prod_{s} \prod_{i \in I_{s,\text{non tronc}}} \frac{\exp(\mathbf{x}'_{(i)}\beta)}{\sum_{j \geq i} \exp(\mathbf{x}'_{(j)}\beta)} \\ \ln LP_{str} &= \sum_{s} \sum_{i \in I_{s,\text{non tronc}}} \left(\mathbf{x}'_{(i)}\beta - \ln \big(\sum_{j \geq i} \exp(\mathbf{x}'_{(j)}\beta) \big) \right) . \end{aligned}$$

Elle diffère de celle sans stratification (eq. 5 page 16) et donne donc lieu à des estimations différentes des coefficients β.

A titre d'illustration, considérons l'estimation de notre modèle avec une stratification selon le genre. On doit alors évidemment enlever la variable *sex* de la liste des covariables. La figure 22 montre les résultats que l'on obtient avec la syntaxe

```
COXREG
   tfp  /STATUS=des(1)
   /STRATA=sex         /* stratification selon le genre */
   /CONTRAST (cohorte)=Indicator
   /METHOD=ENTER edu pnoj cohorte    /* on a enlevé sex */
   /PLOT SURVIVAL.
```

On remarque que par rapport au modèle sans stratification (figure 11 page 21), on gagne un degré de liberté puisque l'on estime un paramètre de moins, celui de *sex*. Il est naturel alors que les statistiques du khi-deux soient

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
4495.008	23.597	4	.000	24.125	4	.000	24.125	4	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 4519.134

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
edu	.022	.022	1.030	1	.310	1.022
pnoj	-.077	.035	4.890	1	.027	.926
cohorte			16.888	2	.000	
cohorte(1)	-.408	.119	11.735	1	.001	.665
cohorte(2)	-.001	.123	.000	1	.994	.999

FIG. 22: Estimation du modèle avec stratification selon le genre

plus faibles. Elles restent cependant toujours aussi significatives. Notons que le modèle de référence pour ces statistiques est ici le modèle naïf stratifié dont le -2LogLik vaut 4519.13 contre 5169.14 sans stratification. Ces valeurs ne sont pas directement comparables.

Les coefficients des covariables autres que *sex* restent très proches de ceux du modèle de la figure 10, page 20, où *sex* est traité comme une covariable. Nous laissons le lecteur vérifier qu'ils diffèrent légèrement plus de ceux du modèle qui ne prend pas en compte *sex*, ni comme variable de stratification, ni comme covariable.

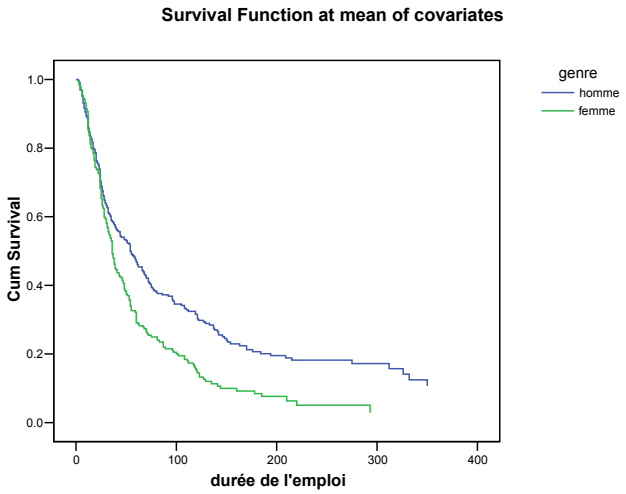

FIG. 23: Courbes de survie stratifiées selon le genre

Bien que l'on estime un seul jeu de coefficients β pour les deux strates, les fonctions de survie $S_s(t, \mathbf{x})$ sont, elles, différentes puisque l'on postule des fonction risque $h_{s,0}(t)$ de référence propres à chaque strate s. La sous-commande `PLOT` dans la syntaxe ci-dessus génère ainsi les deux courbes de la figure 23 où l'on voit que les hommes ont tendance à rester plus longtemps dans un même emploi que les femmes. Cette conclusion va dans le même sens que l'interprétation que l'on peut tirer de la valeur de 1.47 de l'exponentiel du coefficient de *sex* lorsqu'on l'introduit comme covariable. Ce 1.47 nous dit en effet que le risque de terminer un emploi est environ 1.5 fois plus important pour une femme (*sex*=2) que pour un homme (*sex*=1).

9 Modèle avec attributs variant dans le temps

Jusqu'ici, nous avons implicitement admis que les caractéristiques individuelles, les valeurs $\mathbf{x}$ des covariables, sont définies pour chaque cas au début de l'épisode et restent constantes durant tout l'épisode. Ceci est évidemment restrictif est ne permet pas par exemple de prendre en compte une variable telle que le statut marital qui est susceptible de changer pendant la durée d'un emploi.

De façon générale, on peut distinguer deux approches pour prendre en compte des attributs variant dans le temps. L'une, qui n'est malheureusement pas praticable avec SPSS, consiste à partitionner un épisode en sous-épisodes sur lesquels les attributs restent constants. Par exemple, dans nos données l'individu $i = 31$ se marie au 14ème mois d'un emploi qu'il occupe pendant 37 mois. On représente cet emploi par les deux sous épisodes

id	début	fin	marié	fin d'emploi	$\cdots$
31	0	13	0	0	$\cdots$
31	14	37	1	1	$\cdots$

Durant les 13 premiers mois, l'individu participe à la population exposée en tant que non marié et entre 14 et 37 mois en tant que marié. Cette solution qui peut notamment être utilisée dans Stata, S-Plus, R ou TDA, n'est pas applicable dans SPSS qui ne permet pas de préciser le début de l'épisode.

SPSS utilise une seconde méthode spécifique au modèle de Cox. Elle consiste[4] à réévaluer les termes qui entrent dans la vraisemblance partielle

[4]Formellement, on remplace, pour chaque $t_{(i)}$ observé, les profils $\mathbf{x}_{(j)}$ par les profils $\mathbf{x}_{(j)}(t_{(i)})$ dans la formule de vraisemblance partielle. Celle-ci s'écrit alors

$$LP(\beta) \quad = \quad \prod_{i \in I_{\text{non tronc}}} \frac{\exp\left(\mathbf{x}_{(i)}(t_{(i)})'\beta\right)}{\sum_{j \geq i} \exp\left(\mathbf{x}_{(j)}(t_{(i)})'\beta\right)} \ .$$

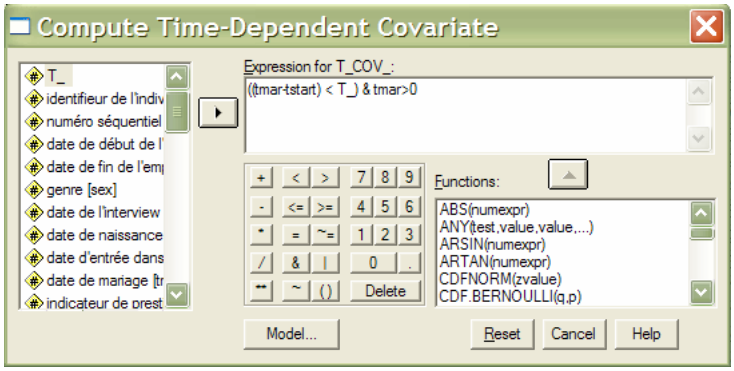

FIG. 24: Défintion de la fonction $x(t)$

(eq. 5 page 16) en chaque $t_{(i)}$ avec les valeurs des covariables valables à cet instant. Ceci suppose que l'on définisse la valeur de l'attribut x variant avec t par une fonction $x(t)$ de la durée t.

Pratiquement, on définit cette fonction dans le dialogue de la figure 24 que l'on active en choisissant 'Cox W/ Time-Dep Cov...' dans le menu de la figure 1. La durée est représentée par une variable interne nommée `T_`. L'expression "`((tmar-tstart) < T_) & tmar>0`" introduite dans le dialogue est une conjonction de deux conditions qui prend la valeur 1 lorsqu'elle est vraie et 0 sinon. La première partie est vraie lorsque le temps entre le début de l'emploi, `tstart`, et le mariage, `tmar`, est inférieur à la durée écoulée `T_`. La seconde condition est vraie lorsqu'il y a eu mariage, c'est-à-dire lorsque `tmar` est non nul. Lorsque les deux conditions sont vérifiées, la fonction assigne la valeur 1, et 0 sinon, à une variable temporaire `T_COV_`. En cliquant sur 'Model...', on active ensuite le dialogue de définition du modèle. La différence est simplement que la liste des variables disponibles dans la base contient cette fois la variable `T_COV_` à valeur dépendant de la durée que l'on vient de définir. Pour l'inclure dans le modèle, il suffit alors de l'ajouter à la liste des covariables.

Par la syntaxe, la fonction $x(t)$ doit être définie entre une instruction `TIME PROGRAM` et l'appel à `COXREG`. La syntaxe est plus souple que le dialogue, car elle permet

- d'assigner le nom de notre choix à la variable créée,
- de définir la fonction $x(t)$ avec plus d'une instruction,
- de définir plusieurs variables à valeurs dépendantes de la durée.

La syntaxe ci-dessous, par exemple, définit une variable `T_mar_` équivalente à la variable `T_COV_` définie dans le dialogue, et estime le modèle de Cox en ajoutant ce nouvel attribut aux facteurs considérés précédemment.

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
5121.749	47.934	6	.000	47.391	6	.000	47.391	6	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 5169.140

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	.386	.097	15.899	1	.000	1.471
edu	.022	.022	1.063	1	.303	1.023
pnoj	-.054	.037	2.155	1	.142	.947
cohorte			18.120	2	.000	
cohorte(1)	-.438	.119	13.535	1	.000	.646
cohorte(2)	-.030	.122	.059	1	.807	.971
T_mar_	-.180	.108	2.775	1	.096	.835

FIG. 25: Modèle avec variable “marié” (*T_mar_*) dépendant du temps

```
COMPUTE ttmar=tmar-tstart.
IF tmar=0  ttmar=10000.
TIME PROGRAM.
   COMPUTE T_mar_ = 0.
   IF T_> ttmar T_mar_ = 1.
COXREG
   tfp  /STATUS=des(1)
   /METHOD=ENTER sex edu pnoj cohorte T_mar_.
```

La figure 25 montre les résultats obtenus pour ce modèle. En particulier on relève que le statut marital (`T_mar_`) n’a pas, au vu de la statistique de Wald, un effet clairement significatif (Sig= 9.6%). On peut se demander si ceci n’est pas la conséquence de deux effets contradictoires, à savoir que les femmes tendent à cesser leur emploi lors du mariage et que les hommes, au contraire, devant la responsabilité du mariage, chercheraient à stabiliser leur emploi. On vérifie cet effet en ajustant des modèles séparés pour les hommes et les femmes. On peut pour cela utiliser la commande `FILTER` comme illustré dans la synatxe ci-dessous

```
COMPUTE women = sex-1.
COMPUTE men = 1 - women.

FILTER OFF.
FILTER BY women.
```

```
TIME PROGRAM.
    COMPUTE T_mar_ = ((tmar-tstart) < T_) & tmar>0.
COXREG
    tfp  /STATUS=des(1)
    /CONTRAST (cohorte) = Indicator
    /METHOD=ENTER edu pnoj cohorte T_mar_.

FILTER OFF.
FILTER BY men.
TIME PROGRAM.
    COMPUTE T_mar_ = ((tmar-tstart) < T_) & tmar>0.
COXREG
    tfp  /STATUS=des(1)
    /CONTRAST (cohorte) = Indicator
    /METHOD=ENTER edu pnoj cohorte T_mar_.

FILTER OFF.
```

Les résultats sont donnés dans les figures 26 et 27. On observe qu'en effet, pour chacun des sexes, l'effet du mariage devient significatif. Comme pressenti, le mariage accroît le risque de fin d'emploi pour les femmes et le diminue pour les hommes.

10 Test de l'hypothèse de proportionnalité

Le modèle semi paramétrique de Cox est très général puisqu'il ne suppose aucune hypothèse sur la distribution des durées des épisodes. Il suppose cependant la proportionnalité des risques. Il convient donc de vérifier que cette hypothèse est raisonnable. Une première approche consiste à vérifier la proportionnalité graphiquement. Une seconde repose sur des tests statistiques. Les tests graphiques examinent les effets, c'est-à-dire les covariables introduites dans le modèle, individuellement. Les tests statistiques portent également sur les effets individuels. Des tests simultanés sur plusieurs covariables sont cependant possibles.

10.1 Tests graphiques

Une première évaluation graphique exploite une propriété des courbes $\ln(-\ln S)$, où S est la fonction de survie. Ces courbes sont appelées LML pour 'Log Minus Log'. Sous l'hypothèse de proportionnalité, on a en effet l'équivalence

$$S(t,\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\beta)} \quad \Leftrightarrow \quad \ln\big(-\ln S(t,\mathbf{x})\big) = \ln\big(-\ln S_0(t)\big) + \mathbf{x}'\beta \ .$$

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1987.589	10.311	5	.067	10.279	5	.068	10.279	5	.068

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 1997.868

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
edu	.030	.032	.877	1	.349	1.031
pnoj	-.102	.062	2.703	1	.100	.903
cohorte			4.900	2	.086	
cohorte(1)	-.022	.186	.014	1	.907	.979
cohorte(2)	.305	.188	2.644	1	.104	1.357
T_mar_	.311	.156	3.979	1	.046	1.365

FIG. 26: Modèle avec variable "marié", femmes

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
2475.847	45.156	5	.000	45.419	5	.000	45.419	5	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 2521.266

b. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
edu	.034	.028	1.418	1	.234	1.034
pnoj	-.023	.046	.262	1	.609	.977
cohorte			24.985	2	.000	
cohorte(1)	-.798	.161	24.503	1	.000	.450
cohorte(2)	-.306	.168	3.343	1	.068	.736
T_mar_	-.643	.152	17.986	1	.000	.526

FIG. 27: Modèle avec variable "marié", hommes

Selon la partie de droite, les courbes LML sont parallèles, en ce sens que pour deux profils $\mathbf{x}_1$ et $\mathbf{x}_2$, la différence entre les courbes LML est, en tout t, égale à $(\mathbf{x}_2 - \mathbf{x}_1)'\beta$ et donc indépendante du temps.

Le test graphique consiste alors à tracer les courbes LML pour chaque valeur d'une covariable, les autres étant maintenues constantes, et de vérifier si cette translation se vérifie. Par exemple, pour le genre on obtient les deux courbes du graphique de gauche dans la figure 28. Il y a proportionnalité si on peut obtenir la courbe du haut par simple glissement de celle du bas vers

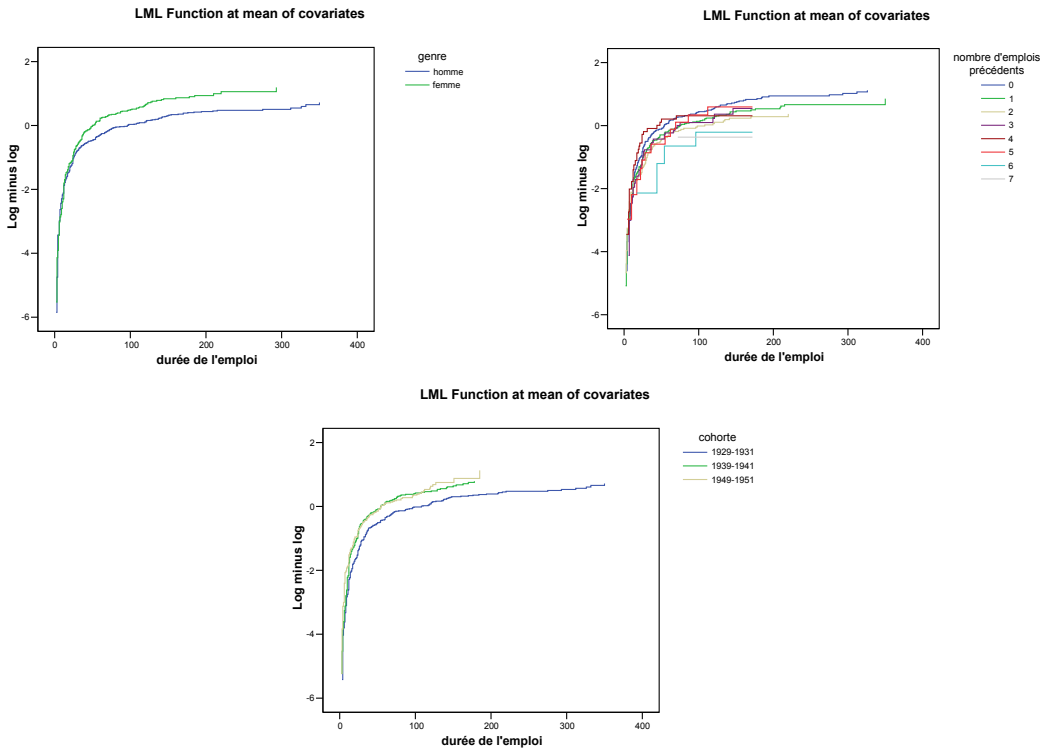

FIG. 28: Comparaison des courbes $\log(-\log S)$

le haut. Pour le sexe, ceci n'est pas tout à fait le cas. La proportionnalité n'est pas strictement vérifiée, mais on en est pas très loin.

Le graphique du bas de la figure 28 montre les courbes LML associées aux trois catégories de *cohorte*. Le parallélisme semble mieux marqué que

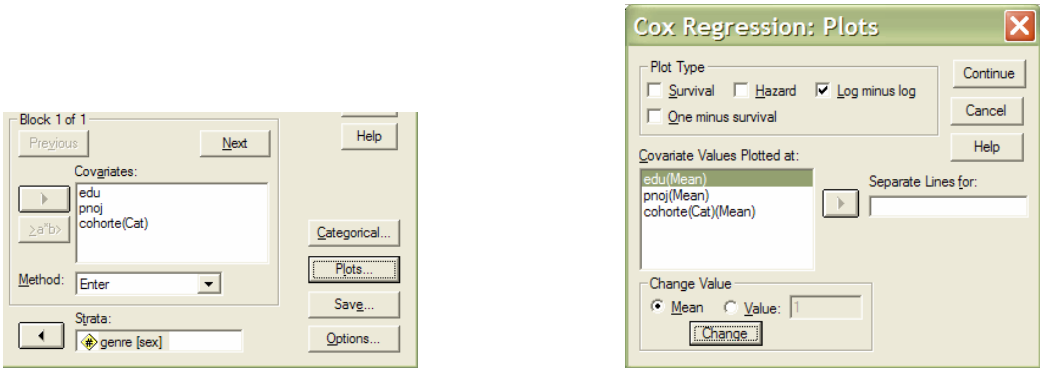

FIG. 29: Demande du plot des courbes LML pour le genre

pour le genre. Notons que les deux courbes du haut qui correspondent aux cohortes 2 et 3 se confondent presque, ce qui découle de l'absence de différence significative déjà constatée entre ces deux cohortes.

Lorsque la variable prend un grand nombre de valeurs, le graphique devient confus et il devient difficile d'en tirer des conclusions. C'est le cas pour la variable *pnoj* dont on peut voir les courbes LML sur le graphique de droite. Ici, la difficulté est encore accrue par le fait que certaines valeurs ne sont observées que pour un petit nombre d'épisodes.

On génère les courbes en utilisant la variable concernée comme variable de stratification et en demandant le plot des courbes LML. Voir le dialogue de la figure 29 ou le code ci-après

```
COXREG
   tfp  /STATUS=des(1)
   /STRATA=sex
   /METHOD=ENTER
   /PLOT=LML.
```

Notons que par rapport au dialogue, la syntaxe proposée utilise un modèle simplifié sans facteurs explicatifs. Les courbes LML générées sont cependant les mêmes.

10.2 Examen des résidus partiels

L'analyse graphique des résidus partiels de Schoenfeld discutés au paragraphe 6.4 constitue une alternative mieux à même de mettre en évidence les situations de non-proportionnalité des risques. L'idée est qu'en cas de proportionnalité des risques, l'écart entre le profil d'un cas i et le profil moyen des cas exposés en t_i devrait être aléatoire et indépendant de t_i. On ne devrait donc pas observer d'effets systématiques dans l'évolution des résidus partiels avec la durée.

Pour examiner cette évolution, deux types de graphiques peuvent être générés à partir des informations que produit `COXREG`.

1. Le diagramme de dispersion des résidus partiels[5] selon la durée t, augmenté de la droite de régression des résidus sur t. Plutôt que t, on retient généralement une transformation $g(t)$ de t tel que par exemple le logarithme $g(t) = \ln(t)$.
2. Le graphique de la séquence des résidus partiels cumulés.[6]

[5] En toute rigueur (Grambsch et Therneau, 1994), il faudrait représenter les résidus partiels recalés (scaled), mais SPSS ne les produit pas.

[6] Il serait plus naturel de représenter l'évolution des résidus cumulés en fonction de t ou

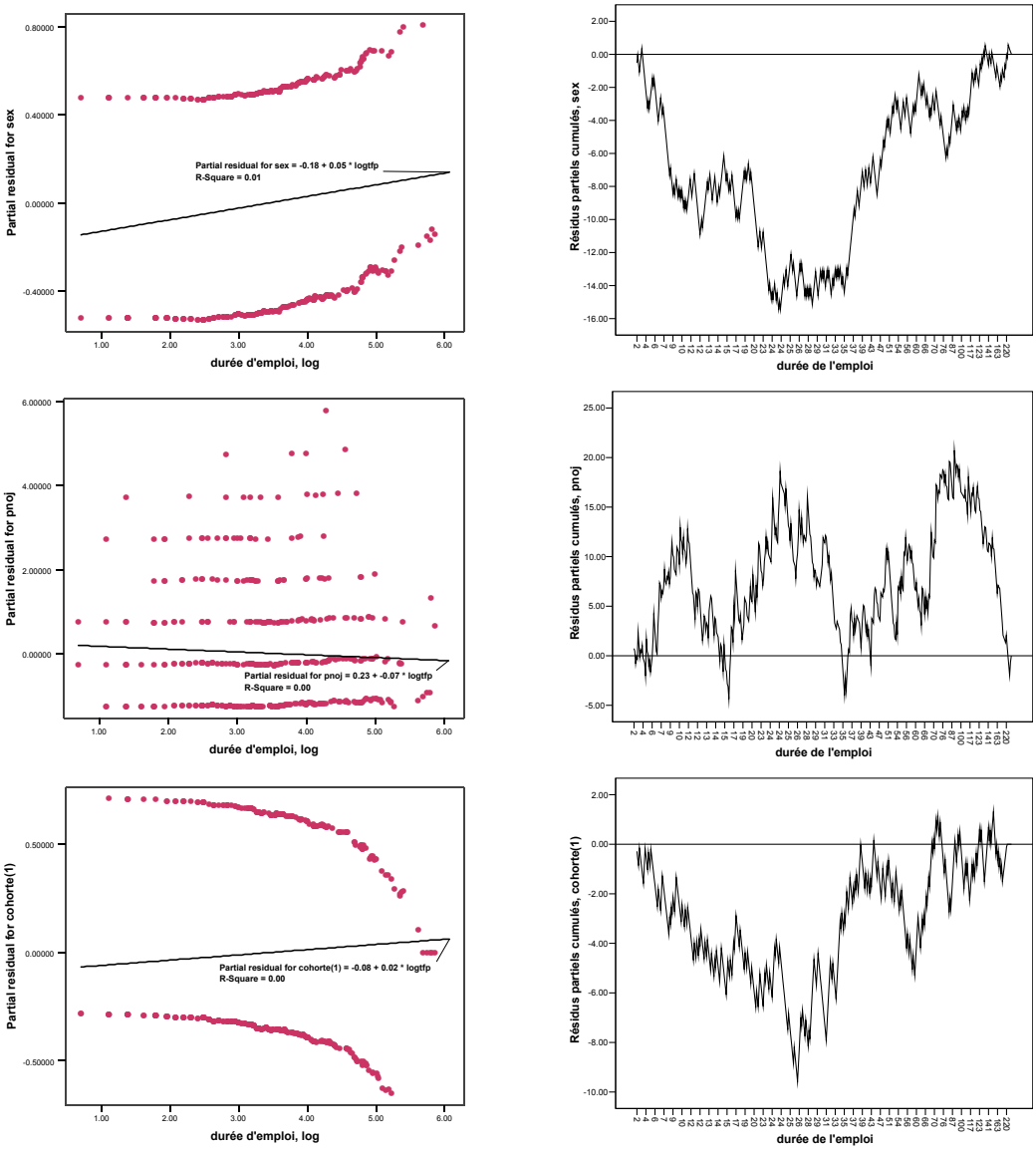

FIG. 30: Résidus partiels (à gauche) et partiels cumulés (à droite), en haut pour *sex*, au milieu pour *pnoj* et en bas pour *cohorte*(1)

La figure 30 montre les graphiques générés pour les variables significatives *sex*, *pnoj* et *cohorte*(1) du modèle de la figure 10 page 20. Le graphique de gauche pour la variable *sex* a été obtenu avec la syntaxe suivante,

---

de $\ln(t)$ plutôt que de leur rang. SPSS ne permet pas cependant de connecter les points successifs dans les diagrammes de dispersion, ce qui rend difficile la lecture des trajectoires d'évolution.

après avoir évidemment généré la série des résidus partiels de Schoenfeld `sch_sex = PR1_1`.

```
IGRAPH
  /VIEWNAME='Scatterplot'
  /X1 = VAR(logtfp) TYPE = SCALE
  /Y  = VAR(sch_sex)  TYPE = SCALE
  /COORDINATE = VERTICAL
  /X1LENGTH=5.0
  /YLENGTH =4.0
  /CHARTLOOK='NONE'
  /SCATTER COINCIDENT = NONE
  /FITLINE METHOD=REGRESSION LINEAR LINE=TOTAL.
```

Le calcul des résidus cumulés et leur présentation graphique ont été réalisés, pour *sex*, avec le code :

```
** Ordonner les cas non tronqués par durée tfp croissante .
COMPUTE tfptmp = 99999. /* cas tronqués     */
IF des=1 tfptmp = tfp.  /* cas non tronqués */
SORT CASES BY tfptmp(A).
** Calcul des résidus cumulés.
COMPUTE tmp = sch_sex.  /* sch_sex résidu pour sex */
FILTER BY tmp.          /* cas non tronqués seulement */
COMPUTE cumPR = tmp.
IF ~missing(LAG(cumPR)) cumPR = cumPR + LAG(cumPR).
VARIABLE LABEL cumPR 'Résidus partiels cumulés, sex'.

** Sequence Charts .
TSPLOT VARIABLES= cumPR
   /ID= tfp
   /NOLOG
   /FORMAT NOFILL REFERENCE(0).
FILTER OFF.
```

Les graphiques de gauche décrivent l'évolution des résidus en fonction du logarithme de la durée $\ln(t)$. La droite de régression indique la tendance. S'agissant du genre, les résidus sont positifs pour les femmes ($sex = 2$) et négatifs pour les hommes. La droite de pente positive, indique que les résidus ont tendance à être négatifs (hommes) pour les petites valeurs de $\ln(t)$ et à devenir positifs (femmes) pour les grandes valeurs. Elle indique une croissance de l'effet genre avec $\ln(t)$. Comme *sex* prend une valeur plus élevée pour les femmes, cet effet s'interprète comme un effet femme. De même, la pente positive pour l'indicatrice de la première cohorte reflète une légère sous-représentation des membres de la première cohorte pour les petites valeurs

de $\ln(t)$, et donc un effet de la première cohorte qui tend à augmenter avec $\ln(t)$. La tendance est toutefois moins marquée que pour le genre. Pour le nombre d'emplois précédents, la faible pente négative indique que l'effet de *pnoj* tend à diminuer faiblement avec la durée.

Les graphiques de droite où l'on voit la séquence des résidus partiels cumulés, sont le résultat du processus où l'on ajoute à chaque pas le nouveau résidu. Dans les trois cas, mais particulièrement pour le genre, *sex*, la courbe reste systématiquement du même côté de la droite horizontale en 0. Ceci semble contredire l'hypothèse de marche au hasard et confirme ainsi les tendances relevées. Les graphiques tendent donc à remettre en cause l'hypothèse de proportionnalité.

10.3 Tests statistiques

Les graphiques précédents permettent de mettre en lumière certains effets non invariants avec t. Sont-ils pour autant statistiquement significatifs ? Nous présentons ici deux tests statistiques de l'hypothèse de proportionnalité que l'on peut effectuer dans SPSS.

Le premier consiste à tester si la pente de la droite de régression des résidus partiels sur la durée t ou une transformation du type $\ln(t)$ est statistiquement significative.[7] Le test consiste donc à vérifier la significativité de la pente des droites représentées sur les graphiques de gauche à la figure 30.

Ce test se fait par le biais de simples régressions linéaires

```
REGRESSION
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS R ANOVA
   /DEPENDENT sch_sex
   /METHOD=ENTER logtfp  .
REGRESSION
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS R ANOVA
   /DEPENDENT sch_pnoj
   /METHOD=ENTER logtfp  .
REGRESSION
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS R ANOVA
   /DEPENDENT sch_coh1
   /METHOD=ENTER logtfp  .
```

[7] En toute rigueur on devrait régresser les résidus partiels recalés que SPSS ne fournit malheureusement pas.

Le tableau 2 récapitule les résultats que l'on obtient pour les trois variables considérées. Il apparaît que la pente est significative pour le résidu partiel associé à *sex*. Les deux autres ne le sont pas. Ceci suggère donc une remise en cause de la constance du rapport de risques entre hommes et femmes seulement.

TAB. 2: Pente de la régression des résidus partiels sur $\ln(t)$

résidus partiels	pente	ratio t	Sig.
sex	.053	2.25	.025
pnoj	-.068	-1.01	.314
cohorte(1)	.024	1.09	.227

La seconde approche consiste à tester une spécification de coefficients variant avec t. On le réalise en complétant le modèle de Cox avec les interactions entre les covariables et t, ou de préférence entre les covariables et $\ln(t)$, et en testant la significativité de ces interactions. Si l'on considère la variable *sex* par exemple, son effet est ainsi mesuré par

$$\beta_{sex} sex + \beta_{sex_lnt} \ln(t)\, sex \quad = \quad \underbrace{\left(\beta_{sex} + \beta_{sex_lnt} \ln(t)\right)}_{\beta_{sex}(t)} sex \tag{17}$$

Il apparaît donc que l'introduction de l'interaction avec $\ln(t)$ n'est qu'une façon de spécifier un coefficient $\beta(t)$ qui varie avec t.

Les interactions doivent être spécifiées entre l'instruction `TIME PROGRAM` et la commande `COXREG`.

```
TIME PROGRAM.
   COMPUTE sex_lnT  = sex*LN(T_).
   COMPUTE edu_lnT  = edu*LN(T_).
   COMPUTE pnoj_lnT = pnoj*LN(T_).
   COMPUTE coh1_lnT = coho1*LN(T_).
   COMPUTE coh2_lnT = coho2*LN(T_).
COXREG
   tfp  /STATUS=des(1)
   /CONTRAST (cohorte)=Indicator
   /METHOD=ENTER sex edu pnoj cohorte
   /METHOD=ENTER sex_lnT edu_lnT pnoj_lnT coh1_lnT coh2_lnT.
```

En introduisant les interactions sous forme d'un bloc, on obtient la valeur de la statistique du rapport de vraisemblance

$$G^2(\text{pas d'interaction} \mid \text{toutes les interactions}),$$

qui compare le modèle avec interactions au modèle sans interactions. Cette statistique fournit un test simultané global de la constance des coefficients. Elle vaut pour notre exemple 8.16 pour 5 degrés de liberté, ce qui est non significatif (Sig= .15) et tend donc à conforter l'hypothèse de proportionnalité.

Les statistiques de Wald que donne directement SPSS pour chaque interaction permettent de tester l'invariance des effets propres à chaque variable. Le tableau 3 résume les résultats obtenus avec la syntaxe ci-dessus. A ce niveau individuel, on constate à nouveau que seul l'effet du sexe varie significativement avec le logarithme de la durée.

TAB. 3: Coefficients des interactions avec $\ln(t)$

covariable x	coefficient de $x \ln(t)$	Wald	Sig.
sex	.225	5.029	.025
edu	.000	.000	.991
pnoj	-.034	.915	.339
coho1	.189	1.979	.160
coho2	.078	.342	.559

Au total, l'hypothèse de proportionnalité semble raisonnable dans cet exemple. Le genre est la seule covariable dont l'invariance avec la durée semble pouvoir être mise en cause.

Pour terminer, notons que les tests discutés permettent de déceler des non proportionnalités qui résultent d'effet qui évoluent linéairement avec t ou une transformation comme le $\ln(t)$. Concrètement, ceci signifie que certains effets non linéaires pourraient échapper à ces tests. Il convient donc de rester prudent avec les conclusions qu'on peut en tirer.

10.4 Quelques solutions en cas de rejet

En cas de rejet de l'hypothèse de proportionnalité, deux solutions relativement simples peuvent être retenues.

La première consiste à modéliser l'évolution des coefficients β dont les tests nous amènent à rejeter la constance. Pratiquement, il s'agit simplement de retenir le modèle avec les interactions entre les covariables concernées et une fonction adéquate de la durée t, $\ln(t)$ par exemple. On a vu en effet (cf eq. 17) que cette façon de faire conduit en fait à spécifier un coefficient variant avec t.

La seconde possibilité est le recours à un modèle stratifié. Comme expliqué à la section 8, en stratifiant on autorise des évolutions du risque, *hazard*,

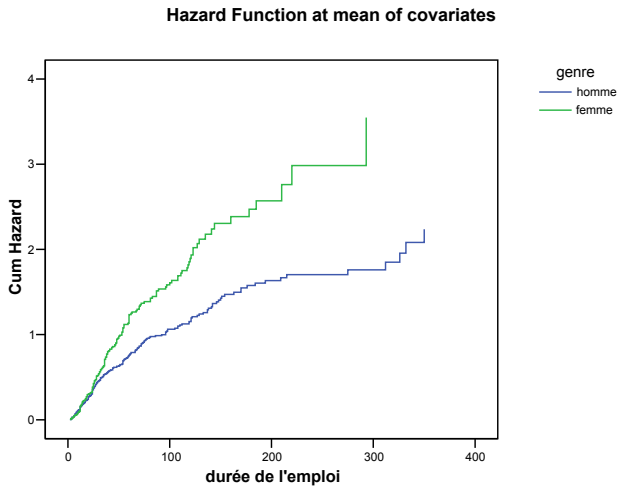

FIG. 31: Fonction de risque cumulé, modèle stratifié selon *sex*

indépendantes dans chacune des strates. Pour notre modèle, où l'on a repéré un effet femme qui augmente avec la durée, il peut ainsi être intéressant de stratifié selon la variable *sex*. On obtient alors le modèle de la figure 22 page 40. En ajoutant `HAZARD` sur la sous-commande `PLOT` de la syntaxe donnée à la page 39, on génère la figure 31 qui montre clairement qu'après environ 2 ans (24 mois), le risque cumulé croît plus rapidement pour les femmes que pour les hommes. Cette approche est plus souple, mais ne permet plus alors de quantifier l'effet de la variable sur le risque. Par ailleurs, la stratification reste limitée aux variables qui n'admettent qu'un petit nombre de valeurs.

11 Pour en savoir plus

Le modèle de Cox est devenu incontournable et tous les ouvrages consacrés à l'analyse biographique, à la *Event History Analysis* ou encore à l'analyse de la survie, lui consacrent une part plus ou moins importante. On peut consulter par exemple Allison (1984), Blossfeld et Rowher (2002), Courgeau et Lelièvre (1989), Kleinbaum (1996), Yamaguchi (1991). L'ouvrage de Hosmer et Lemeshow (1999) donne une présentation orientée utilisateur très complète avec de nombreuses illustrations issues du milieu médical. Pour les aspects plus statistiques et des développements plus avancés du modèle voir en particulier Harell (2001) et Therneau et Grambsch (2000).

Les procédures SPSS sont bien adaptées pour des analyses simples (événement et calendrier unique, troncature à droite seulement, ...). Les formes plus avancées, incluant en particulier des formulations multi-niveaux comme

la *shared frailty*, nécessitent cependant des logiciels plus adaptés tels que aML, Stata, S-Plus, R ou TDA, les deux derniers étant des logiciels gratuits. TDA est particulièrement bien adapté pour traiter les épisodes et événements multiples. On trouve par ailleurs de nombreux programmes pour S-Plus ou R dans les ouvrages avancés tels que Harell (2001) et Therneau et Grambsch (2000).

Pour terminer, mentionnons encore quelques sites internet sur l'analyse de données longitudinales où l'on trouvera des informations utiles :

http://www2.unil.ch/pavie/documentation/introduction.htm
: Les notes, dont ce guide, sur la mise en œuvre de l'analyse des biographies avec SPSS réunies par Jean-Marie Le Goff.

http://www.xs4all.nl/~jhckx/
: Le site de John Hendrickx qui propose des syntaxes SPSS pour reproduire certains exemples de Blossfeld et Rowher (2002).

http://www.iser.essex.ac.uk/teaching/stephenj/ec968/index.php
: Le site de Stephen P. Jenkins de l'Université d'Essex où l'on trouve les notes de son excellent cours sur la *Survival analysis* avec Stata, ainsi que des jeux de données et des syntaxes Stata (do files).

http://www.lrz-muenchen.de/~wlm/tdaframe.htm
: Le site de Wolfgang Ludwig-Mayerhofer qui propose un tutoriel sur l'*Event history analysis* avec TDA. Le tutoriel utilise un jeu de données tiré de Yamaguchi (1991).

A Annexes

A.1 Les données : contenu du fichier rrdat.1

Le fichier de donnée `rrdat.1` que l'on trouve sur le site de TDA est un fichier texte dont les premières lignes sont :

```
1  1 555 982  1 982 351 555 679  34  -1 17
2  1 593 638  2 982 357 593 762  22  46 10
2  2 639 672  2 982 357 593 762  46  46 10
2  3 673 892  2 982 357 593 762  46  -1 10
3  1 688 699  2 982 473 688 870  41  41 11
3  2 700 729  2 982 473 688 870  41  44 11
3  3 730 741  2 982 473 688 870  44  44 11
3  4 742 816  2 982 473 688 870  44  44 11
3  5 817 828  2 982 473 688 870  44  -1 11
4  1 872 926  2 982 604 872 872  55  -1 13
.  .   .   .  .   .   .   .   .   .   .  .
```

Le fichiers contient en tout 600 lignes correspondant aux 600 épisodes, les colonnes correspondent à 12 variables qui sont dans l'ordre :

1	id	identifieur de l'individu
2	noj	numéro séquentiel de l'emploi
3	tstart	date de début de l'emploi
4	tfin	date de fin de l'emploi
5	sex	genre (1 homme, 2 femme)
6	ti	date de l'interview
7	tb	date de naissance
8	te	date d'entrée dans sur le marché de l'emploi
9	tmar	date de mariage (0 si pas de mariage)
10	pres	indicateur de prestige de l'emploi courant
11	presn	indicateur de prestige de l'emploi suivant (-1 si non applicable)
12	edu	niveau d'éducation à l'entrée sur le marché de l'emploi

Toutes les dates (tstart, tfin, ti, tb, te, tmar) sont en nombre de mois depuis janvier 1900. Le premier emploi considéré débute ainsi le 555ème mois qui correspond à mars 1946.

La syntaxe ci-dessous, que l'on trouve également parmi les syntaxes fournies dans le fichier `rrdat_create.sps`, génère le fichier SPSS documenté et augmenté des variables construites

13	tfp	durée de l'emploi
14	logtfp	logarithme de la durée de l'emploi
15	des	troncature (1 si fin d'emploi, 0 sinon)
16	cohorte	cohorte (1='1929-1931', 2='1939-1941', 3='1949-1951')
17	coho1	cohorte 1 (1 si cohorte 1929-1931, 0 sinon)
18	coho2	cohorte 2 (1 si cohorte 1939-1941, 0 sinon)
19	coho3	cohorte 3 (1 si cohorte 1949-1951, 0 sinon)
20	lfx	expérience sur marché du travail
21	pnoj	nombre d'emplois précédents

TAB. 4: Syntaxe de création du fichier `rrdat.sav`

```
* Lecture du fichier original rrdat.1 (fichier ASCII) tel que fourni
* avec TDA sur http://steinhaus.stat.ruhr-uni-bochum.de/tda.html .

* Lecture du fichier et création des noms de variables.
DATA LIST FILE='./rrdat.1' FREE
 / id noj tstart tfin sex ti tb te tmar pres presn edu.

* Attribution d'étiquettes aux variables.
VARIABLE LABELS
    id     'identifieur de l''individu'
    noj    'numéro séquentiel de l''emploi'
    tstart 'date de début de l''emploi'
    tfin   'date de fin de l''emploi'
    sex    'genre'
    ti     'date de l''interview'
    tb     'date de naissance'
    te     'date d''entrée sur le marché de l''emploi'
    tmar   'date de mariage'
    pres   'indicateur de prestige de l''emploi courant'
    presn  'indicateur de prestige de l''emploi suivant'
    edu    'niveau d''éducation à l''entrée sur le marché d''emploi'
    .
* Etiquettes de valeurs pour variables catégorielles.
VALUE LABELS
    sex  1 'homme' 2 'femme'
   /tmar 0 'pas de mariage'.
```

```
* La valeur -1 de presn sera traitée comme valeur manquante.
MISSING VALUES presn(-1).

* Toutes les valeurs sont entières, on supprime les décimales.
FORMAT all(F8.0).

*** Construction de variables.

* durée d'emploi.
COMPUTE tfp = tfin - tstart + 1.
COMPUTE logtfp = ln(tfp).
VARIABLE LABELS
   tfp     'durée de l''emploi'
  /logtfp  'durée d''emploi, log'.

* variable de troncature.
COMPUTE des = (tfin < ti).
VARIABLE LABELS
    des    'destination (1 si fin emploi, 0 sinon)'.
VALUE LABELS
    des 0  'tronqué' 1 'fin emploi'.

* cohorte.
IF tb > 348 and tb < 385 cohorte = 1.
IF tb > 468 and tb < 505 cohorte = 2.
IF tb > 588 and tb < 625 cohorte = 3.
VALUE LABELS
   cohorte 1 '1929-1931' 2 '1939-1941' 3 '1949-1951'.

* recodage de cohorte en variables indicatrices.

RECODE cohorte (1=1) (else=0) INTO coho1.
RECODE cohorte (2=1) (else=0) INTO coho2.
RECODE cohorte (3=1) (else=0) INTO coho3.
VARIABLE LABELS
     coho1      '1929-1931'
    /coho2      '1939-1941'
    /coho3      '1949-1951'.
```

```
* Expérience sur le marché du travail.
COMPUTE lfx = tstart - te.
VARIABLE LABELS
    lfx    'expérience sur marché du travail'.

* Nombre de jobs précédents.
COMPUTE pnoj = noj - 1.
VARIABLE LABELS
    pnoj    'nombre d''emplois précédents'.

FORMAT
    tfp lfx pnoj (F8.0)
    des cohorte coho1 coho2 coho3 (F3.0).

*** Quelques résumés statistiques des données.

DESCRIPTIVES VARIABLES=ALL.
FREQUENCIES VARIABLES=sex cohorte noj edu des.

*** Enregistrement du fichier SPSS.

SAVE OUTFILE='./rrdat.sav'.
```

A.2 Principe de l'estimation de $S_0(t)$

On suppose le risque $h_0(t)$ constant entre deux temps successifs $t_{(i)}$ et $t_{(i+1)}$ observés de survie, ce qui permet d'écrire

$$\begin{aligned}\frac{S(t_{(i+1)},\mathbf{x})}{S(t_{(i)},\mathbf{x})} &= \frac{S_0(t_{(i+1)})^{\exp(\mathbf{x}'\beta)}}{S_0(t_{(i)})^{\exp(\mathbf{x}'\beta)}} \\ &= \left(\frac{S_0(t_{(i+1)})}{S_0(t_{(i)})}\right)^{\exp(\mathbf{x}'\beta)} \\ &= \big(1-h_0(t_{(i)})\big)^{\exp(\mathbf{x}'\beta)} \ .\end{aligned}$$

D'où

$$S_0(t_k) = \prod_{i<k}(1-\hat{h}_0(t_{(i)}) \ .$$

Posons $\alpha_i = 1-h_0(t_{(i)})$ pour alléger un peu les notations. On obtient une estimation $\hat{S}_0(t_k)$ de $S_0(t_k)$ en remplaçant les $h_0(t_i)$ et donc les α_i par des estimations. L'estimateur de S_0 s'écrit ainsi

$$\hat{S}_0(t_{(k)}) = \prod_{i<k}\hat{\alpha}_i = \prod_{i<k}(1-\hat{h}_0(t_{(i)}) \ ,$$

forme dont on peut relever l'analogie avec l'estimateur de Kaplan-Meier. La différence tient ici dans la façon d'estimer les α_i. On les estime par une procédure de maximum de vraisemblance conditionnelle aux estimations $\hat{\beta}$. Les estimations du maximum de vraisemblance des α_i sont les solutions du système

$$\sum_{j\in D_i}\frac{\exp(\mathbf{x}_j'\hat{\beta})}{1-\alpha_i^{\exp(\mathbf{x}_j'\hat{\beta})}} = \sum_{j\in R_i}\exp(\mathbf{x}_j'\hat{\beta})$$

où D_i est l'ensemble des cas dont le temps de survie observé est $t_{(i)}$ et R_i l'ensemble des cas exposés au risque en $t_{(i)}$.

En l'absence d'égalités, ces solutions s'écrivent :

$$\hat{\alpha}_i = \left(1-\frac{\exp(\mathbf{x}_i'\hat{\beta})}{\sum_{j\in R_i}\exp(\mathbf{x}_j'\hat{\beta})}\right)^{\exp(-\mathbf{x}_i'\hat{\beta})} \ .$$

Certains logiciels dont SPSS utilisent l'approximation de Breslow des $\hat{\alpha}_i$, soit, en notant d_i le nombre d'événements en $t_{(i)}$

$$\tilde{\alpha}_i = \exp\Big(\frac{-d_i}{\sum_{j\in R_i}\exp(\mathbf{x}_j'\hat{\beta})}\Big) \ .$$

Références

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institue 50*, 277–290.

Allison, P. D. (1984). *Event history analysis, Regression for longitudinal event data*. QASS 46. Beverly Hills and London: Sage.

Blossfeld, H.-P. et G. Rowher (2002). *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ: Lawrence Erlbaum.

Courgeau, D. et É. Lelièvre (1989). *Analyse démographique des biographies*. Paris: Editions de l'INED.

Cox, D. R. et E. J. Snell (1989). *The Analysis of Binary Data* (2nd ed.). London: Chapman and Hall.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B 34*(2), 187–220.

Grambsch, P. M. et T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*, 515–526.

Harell, Jr., F. E. (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Hosmer, D. W. et S. Lemeshow (1999). *Applied Survival Analysis, Regression Modeling of Time to Event Data*. New York: Wiley.

Kleinbaum, D. G. (1996). *Survival Analysis, a Self-Learning Text*. New York: Springer-Verlag.

Mayer, K. U. et E. Brückner (1989). *Lebensverläufe und Wohlfahrtsentwicklung. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929-1931, 1939-1941, 1949-1951*. Materialen aus der Bildungsforschung. Berlin: Max-Planck Institut für Bildungsforschung.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika 78*(3), 691–692.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC: The American Sociological Association.

Rohwer, G. et U. Pötter (2002). TDA user's manual. Software, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

SPSS Inc. (2003). *SPSS 12 Command Syntax Reference*. Chicago, IL: SPSS Inc.

Therneau, T. M. et P. M. Grambsch (2000). *Modeling Survival Data*. New York: Springer.

Yamaguchi, K. (1991). *Event history analysis*. ASRM 28. Newbury Park and London: Sage.

Index

accelerated failure times, 14
AIC, 26

BIC, 26
Block, 24

coefficient variant avec t, 51
`Covariates`, 3
Cox, 1
`COXREG`
    commande, 4
    `CONTRAST`, 11, 19
    `CRITERIA`, 4, 16, 32
    `METHOD`, 4, 24, 28, 32
    `PATTERN`, 36
    `PLOT`, 36
    `SAVE`, 33, 37
    `STATUS`, 4
    `STRATA`, 39

densité, 14
durée, 13

épisodes tronqués, voir troncature

`FILTER`, 43

`GRAPH`, 38

hazard
    baseline, 15
    cumulé, 14
    rate, 14
    ratio, 5, 10

index, 61
interaction avec $\ln t$, 51

khi-deux
    rapport de vraisemblance, 21, 22
    Score, 17, 21
    Wald, 17, 18, 21

LogLik, log vraisemblance, 21

`MATRIX - END MATRIX`, 38
modèle
    avec variable dépendant de t, 41
    stratifié, 38

output, 5, 6, 11, 12, 20, 21, 29, 31, 43

panels, 1
pseudo R^2
    Cox-Snell, 23
    Nagelkerke, 23

ratio critique, 18
`RECODE`, 9
résidus partiels, 33, 47, 50
risques
    proportionnels, 14
    ratio de, 5, 10, 15
`rrdat.1`, 2, 55
`rrdat.sav`, 2
`rrdat_create.sps`, 55

Schoenfeld, voir résidus partiels
Score Test, voir khi-deux
`Status`, 2, 13
Step, 24
Student, test de, 18
syntaxe, 4

`T_`, 42
`T_COV`, 42
TDA, 2
tests de proportionnalité
    courbes LML, 44
    résidus partiels, 47
    statistiques, 50
`Time`, 2, 13
`TIME PROGRAM`, 42, 51
troncature, 2, 13

variable
    catégorielle, 7–12, 19
    variant avec t, 41
vraisemblance
    partielle, 16, 39, 41
    rapport de, 22

Wald, voir khi-deux