```
Call:
glm(formula = nbinf ~ age + bmi + crp + prealb + sexe, family = quasipoisson❹)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6912  -0.9806  -0.3065   0.5027   1.9639
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.700324   1.442488  -4.645 1.18e-05 ***
age          0.048145   0.015048   3.199  0.00192 **
bmi          0.028300   0.020111   1.407  0.16291
crp          0.008546   0.001602   5.335 7.35e-07 ***
prealb       0.779498   1.453737   0.536  0.59317
sexe         0.830061   0.263852   3.146  0.00226 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
(Dispersion parameter for quasipoisson family taken to be 0.7381137❺)
    Null deviance: 117.701  on 93  degrees of freedom
Residual deviance:  74.608  on 88  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4
```

Dans le cas présent, le paramètre est inférieur à 1. Il n'y a donc pas surdispersion mais plutôt une « sous »-dispersion. Les résultats de la régression de Poisson étaient donc vraisemblablement admissibles.

3.
Modèle de Cox

Le modèle de Cox est un modèle de régression adapté au cas d'une *variable à expliquer censurée* comme la survie, au même titre que le modèle de régression linéaire est adapté au cas d'une variable à expliquer quantitative ou le modèle de régression logistique au cas d'une variable à expliquer binaire.

Ainsi, en pratique, un modèle de Cox permettra de relier la survie d'un patient à une liste de facteurs pronostiques, tout en mettant en évidence le poids spécifique de chacun d'entre eux.

Sur un plan formel, le modèle de Cox postule que le risque instantané de décès d'un patient caractérisé par un ensemble de p variables explicatives Z_1, ..., Z_p (facteurs pronostiques quantitatifs ou qualitatifs binaires, par exemple l'âge, le sexe, la présence de métastases, etc.) peut s'écrire sous la forme :

$$h(t) = h_0(t) \cdot \exp(a_1Z_1 + ... + a_pZ_p)\ (1)$$

où $h_0(t)$ est une fonction quelconque ne dépendant que du temps et a_1, ..., a_p sont des constantes. On notera avec intérêt que l'évolution temporelle du risque de décès n'est pas imposée par le modèle. Les conditions suivantes doivent néanmoins être remplies :

> Le rapport des risques instantanés de décès de deux sujets est indépendant du temps (hypothèse des risques proportionnels). Le logarithme de h(t) est une fonction linéaire des Z_i (hypothèse de loglinéarité).

Nous verrons plus bas comment vérifier de telles hypothèses.

Un tel modèle peut sembler difficile à saisir au premier abord. Il est pourtant d'un usage très simple. En effet :

- il n'y a pas de condition de validité à vérifier quant à la forme que doit avoir la courbe de survie ;
- le test que tout utilisateur potentiel attend, c'est-à-dire : « la survie est-elle spécifiquement liée à la variable Z_i, c'est-à-dire une fois pris en compte l'effet des variables Z_j, $i \neq j$? » correspond au simple test de $a_i = 0$;
- enfin, si Z_i est binaire, $\exp(a_i)$ est égal au risque de décès relatif à l'exposition au facteur Z_i. Nous avions un résultat similaire dans la régression logistique où $\exp(a_i)$ était égal à un odds-ratio.

---

[1] Si S(t) est la fonction de survie associée à h(t) et si $S_0(t)$ est associée à $h_0(t)$, on a de façon équivalente : $\text{Log}[S(t, Z_1, ..., Z_p)] = \text{Log}[S_0(t)] \cdot \exp(a_1Z_1 + ... + a_pZ_p)$.

Le test du log-rank comme cas particulier du modèle de Cox

En quelques mots : nous avons remarqué dans des chapitres précédents qu'une régression linéaire avec pour seule variable explicative une variable qualitative binaire était équivalente à un test t, alors qu'une régression logistique sur une variable explicative binaire était équivalente à un test du chi-2. Un modèle de Cox portant sur une simple variable explicative binaire z_i conduit, lui, à un test du log-rank (2).

En pratique : dans le chapitre sur les données de survie (p. 87), nous avons analysé les données d'une étude portant sur 65 patients atteints de myélome. A l'aide du test du log-rank, nous avons recherché si la survie était significativement différente que l'on ait, ou non, une protéinurie de type Bence-Jone. Voyons maintenant ce que donnerait un modèle de Cox expliquant la survie par une seule variable explicative Z_i : la variable binaire « protéinurie de type Bence-Jone ».

La procédure SAS, PROC PHREG conduit aux résultats :

The PHREG Procedure

Data Set: BF.MYEL
Dependent Variable: TEMPS
Censoring Variable: DECES
Censoring Value(s): 0
Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
65	48	17	26.15

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	309.716	307.678	2.038 with 1 DF (p=0.1534)
Score	.	.	2.010 with 1 DF (p=0.1562) ❺
Wald	.	.	1.984 with 1 DF (p=0.1589)

Analysis of Maximum Likelihood Estimates ❶

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
BENCE_J	1	0.451990 ❷	0.32087 ❸	1.98423	0.1589 ❹	1.571

[2] Plus précisément, s'il n'y a pas de décès *ex aequo*.

Les résultats portant sur le coefficient a_i de la variable « Bence_J » sont en ❶. Le coefficient lui-même est en ❷, son écart type en ❸ et le « p » correspondant au test de l'hypothèse $a_i = 0$ est en ❹ ou en ❺. Il existe, en effet, plusieurs façons de tester l'hypothèse nulle (test de score en ❺, test de Wald, test du maximum de vraisemblance). Ces « p » sont à comparer au résultat p = 0,1562 que nous avions trouvé page 96 pour le test du log-rank.

Le modèle de Cox : un test du log-rank avec possibilité d'ajustement

En quelques mots : si l'on enrichit le modèle ci-dessus, il va maintenant être possible de rechercher une différence de survie entre les groupes « Bence_J = 0 » et « Bence_J = 1 » après ajustement sur plusieurs covariables. Il faut, bien sûr, garder à l'esprit qu'un ajustement est tributaire d'un modèle, et qu'il est donc moins performant que le contrôle expérimental des covariables envisagées.

En pratique : nous allons maintenant estimer un modèle de Cox qui expliquera la survie non seulement par la variable « protéinurie du type Bence-Jone », mais aussi par une liste de facteurs pronostiques possibles tels que : « urée plasmatique » (en fait son logarithme, que l'on notera « log_urée »), « hémoglobinémie » (Hb), « calcémie » (calcium) et « âge » (âge).

Nous obtenons grâce à la procédure SAS, PROC PHREG (3) :

Total	Event	Censored	Percent Censored
65	48	17	26.15

Testing Global Null Hypothesis: BETA=0 ❶

Criterion	ithout Covariates	ith Covariates	Model Chi-Square
-2 LOG L	309.716	291.304	18.412 with 5 DF (p=0.0025)
Score	.	.	20.691 with 5 DF (p=0.0009)
ald	.	.	19.534 with 5 DF (p=0.0015)

Analysis of Maximum Likelihood Estimates ❷

Variable ❸	DF	Parameter Estimate ❹	Standard Error	ald Chi-Square	Pr > Chi-Square ❺	Risk Ratio
BENCE_J	1	0.576023	0.33633	2.93322	0.0868	1.779
LOG_UREE	1	1.919298	0.63784	9.05440	0.0026	6.816
CALCIUM	1	0.169289	0.09996	2.86818	0.0903	1.184
HB	1	-0.130366	0.06230	4.37920	0.0364	0.878
AGE	1	-0.026560	0.01655	2.57551	0.1085	0.974

[3] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

En ❶ est testée l'association globale entre les 5 variables explicatives et la survie (test de peu d'utilité pratique).

Les résultats véritablement intéressants sont introduits en ❷. Les variables sont en ❸, les coefficients a_i en ❹ et le « p » correspondant au test de l'hypothèse $a_i = 0$ en ❺. Les covariables « log_urée » et « Hb » sont significativement liées à la survie, mais malgré l'ajustement, l'association entre « Bence_J » et la survie persiste quant à elle à ne pas être significative, comme l'indique la valeur de p = 0,0868.

Modèle de Cox avec stratification

En quelques mots : dans le chapitre sur la régression logistique, nous avons vu (p. 172) qu'il existait un modèle particulièrement adapté aux enquêtes cas-témoins sur séries stratifiées : la régression logistique conditionnelle. Dans le cadre d'étude portant sur des données de survie, la question de la stratification se pose tout autant.

Dans le chapitre clôturant cette partie consacrée aux modèles multivariés (p. 229) nous reviendrons largement sur les notions de stratification et d'ajustement. Voyons d'ores et déjà comment, en pratique, le modèle de Cox intègre ces deux approches.

En pratique : imaginons que les variables « log_urée » et « calcium » introduites dans le modèle de Cox précédent ne vérifient pas l'hypothèse des risques proportionnels. Une solution serait alors de transformer les deux variables concernées en variables qualitatives, puis de procéder à une stratification plutôt qu'à un ajustement. Le modèle de Cox correspondant opérera ainsi une stratification sur « log_urée » et « calcium » et un ajustement sur « âge » et « Hb ».

En pratique, chaque variable est découpée en deux classes, « log_urée » à la hauteur de 1,4 et « calcium » à la hauteur de 10. Finalement, nous obtenons numériquement (4) :

The PHREG Procedure

Summary of the Number of Event and Censored Values ❶

Stratum	LOG_UREE	CALCIUM	Total	Event	Censored	Percent Censored
1	<1.4	<10	16	11	5	31.25
2	<1.4	>10	24	18	6	25.00
3	>1.4	<10	10	7	3	30.00
4	>1.4	>10	15	12	3	20.00
Total			65	48	17	26.15

[4] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	192.260	186.117	6.142 with 3 DF (p=0.1049)
Score	.	.	6.095 with 3 DF (p=0.1071)
Wald	.	.	5.908 with 3 DF (p=0.1162)

Analysis of Maximum Likelihood Estimates ❷

Variable ❸	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square ❹	Risk Ratio
BENCE_J	1	0.398283	0.35366	1.26830	0.2601	1.489
HB	1	-0.128842	0.06553	3.86598	0.0493	0.879
AGE	1	-0.022146	0.01743	1.61426	0.2039	0.978

En ❶ nous trouvons les détails de la stratification. En ❷ débute la partie concernant l'ajustement. Les variables sont en ❸, les « p » correspondant au test de l'hypothèse nulle $a_i = 0$ en ❹. Ces résultats sont voisins de ceux obtenus sans stratification.

Modèle de Cox avec covariables dépendantes du temps

En quelques mots : dans une étude portant sur la survie ou sur toute autre variable censurée, il n'est pas rare que les facteurs de risque auxquels sont exposés les sujets aient un niveau fluctuant au cours du temps. Ainsi, dans une étude s'intéressant à l'effet d'un régime riche en fibres sur la prévention de la survenue d'un infarctus du myocarde, le régime pourra n'être suivi qu'un certain laps de temps, un tabagisme pourra être arrêté ou une hypertension artérielle pourra, elle, apparaître. On parle alors de covariables dépendantes du temps.

Le modèle de Cox permet de gérer de telles variables. La difficulté vient de la présentation des données qui est différente : il faut, en général, découper le suivi de chaque patient en tronçons sur lesquels les covariables sont constantes. Le fichier à analyser compte alors une ligne par tronçon et non une ligne par patient.

En pratique : prenons l'exemple d'une cohorte de patients alcooliques. La variable à expliquer est la survenue d'un sevrage. Les variables explicatives sont l'âge, le sexe (toutes deux non dépendantes du temps) ainsi que la survenue oui/non d'événements de vie négatifs depuis la dernière consultation. Cette dernière variable est, elle, dépendante du temps.

Cent vingt-huit patients sont ainsi suivis sur une période allant de quelques mois à plusieurs années. Pour analyser les données, il est souvent nécessaire de les présenter de la façon suivante :

NUM	NCS	SEVRE	AGE	SEXE	EDVNEG	t1	t2
101	2	0	53	1	0	1	121
101	3	0	53	1	1	121	274
101	4	0	53	1	0	274	448
...	...	...	...	...	...	...	...
366	3	0	52	2	1	121	274

Il ne s'agit pas du format : un patient (NUM) par ligne. Il ne s'agit pas tout à fait non plus du format : une consultation (NCS) par ligne.

Chaque ligne représente plutôt une période débutant à « t1 » et se terminant à « t2 ». Durant ce laps de temps les variables mesurées ont une valeur constante.

Le logiciel R et sa fonction « coxph » de la librarie « survival » donne les résultats ([5]) :

```
Call:
coxph(formula = Surv(t1, t2, SEVRE) ~ EDVNEG + AGE + SEXE, data = alc)
          coef exp(coef) se(coef)      z     p
EDVNEG -0.50164     0.606  0.30165 -1.663 0.096❶
AGE     0.00238     1.002  0.00838  0.284 0.780
SEXE   -0.34895     0.705  0.22713 -1.536 0.120
Likelihood ratio test=5.6  on 3 df, p=0.133  n= 744
```

Les résultats sont à interpréter comme à l'habitude : il y a bien une association négative entre la survenue d'événements de vie négatifs et l'obtention d'un sevrage, mais le test de l'association n'est pas significatif puisque p = 0,096 en ❶.

On remarquera ici qu'un patient peut être sevré, rechuter et être de nouveau sevré : il s'agit donc possiblement d'événements répétés dans le temps. Une analyse prenant en compte cette particularité est présentée dans le chapitre « Mesures répétées », p. 201.

Modèle de Cox : le prix à payer

En quelques mots : la marche à suivre pour réaliser un diagnostic de régression d'un modèle de Cox est identique à celle que nous avons suivie pour les modèles de régression linéaire ou de régression logistique. En théorie, trois points devraient être abordés : vérification de l'hypothèse des risques proportionnels, vérification de l'hypothèse de loglinéarité et enfin recherche de sujets marginaux par l'examen des résidus.

– La vérification de l'hypothèse des risques proportionnels se fait à l'aide de tests statistiques et de schémas.

– L'hypothèse de loglinéarité peut, en théorie, être étudiée à l'aide de schémas. En réalité ces schémas sont peu informatifs et rarement utilisés.

[5] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

– Enfin, la recherche de sujets « marginaux » ou de sujets « sensibles » sera réalisée à partir de paramètres indiquant l'importance de la variation des coefficients a_i quand un sujet donné est enlevé de l'échantillon.

– Il est rare qu'un seul logiciel propose l'ensemble de ce panel de vérifications...

En pratique : les trois points exposés ci-dessus vont maintenant être abordés à partir de l'exemple étudiant la survie de patients atteints de myélome ([6]). Nous allons plus précisément nous pencher sur la validité du modèle de Cox étudié plus haut, reliant la survie aux variables : « protéinurie type Bence-Jone » (Bence_J), « fonction rénale » (log_urée), « hémoglobinémie » (Hb), « âge » (âge) et enfin « calcémie » (calcium).

Hypothèse des risques proportionnels

Le logiciel R et sa fonction « cox.zph » de la bibliothèque « survival » est très complet pour évaluer l'acceptabilité de l'hypothèse des risques proportionnels.

Nous avons d'une part les schémas de la figure 3.1. Ces graphiques donnent l'allure de l'évolution temporelle de résidus appelés résidus de Shoenfeld. Si l'hypothèse des risques proportionnels est vérifiée, ces résidus ont en théorie un aspect totalement aléatoire et l'évolution temporelle moyenne est une droite horizontale. Si l'hypothèse des risques proportionnels n'est pas vérifiée, par exemple si le facteur de risque est important au début du suivi du patient mais pas à la fin, alors les résidus seront, sur le schéma, négatif puis positif et l'évolution temporelle moyenne sera une courbe croissante (ou le contraire en fonction du codage du facteur de risque).

Nous constatons ici que, pour chaque facteur de risque, l'évolution moyenne des résidus au cours du temps prend l'allure d'une fonction horizontale, constante.

Cette impression peut être objectivée à l'aide d'un test statistique :

```
             rho   chisq      p
BENCE.J   0.0110 0.00577 0.9395
LOG.UREE -0.2087 3.22666 0.0724
CALCIUM  -0.1620 1.86665 0.1719
HB       -0.0173 0.01598 0.8994
AGE      -0.0832 0.40183 0.5261
GLOBAL       NA 6.47111 0.2630
```

Nous constatons qu'aucun « p » n'est inférieur à 5 %. Il semble donc qu'aucune variable explicative ne soit en contradiction flagrante avec l'hypothèse des risques proportionnels.

[6] Les données et les syntaxes sas et R des exemples ci-dessous sont disponibles sur le site Internet du livre.

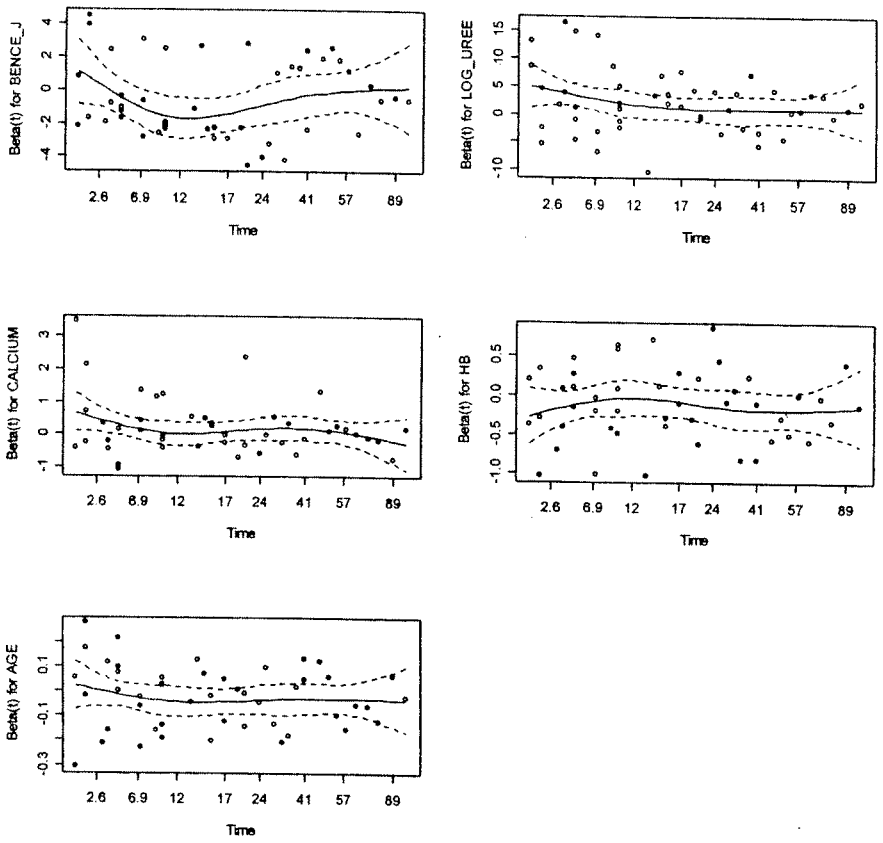

Fig. 3.1 — *Appréciation graphique de l'acceptabilité de l'hypothèse des risques proportionnels dans le modèle de Cox.*

Hypothèse de loglinéarité

Une étude des résidus peut permettre, dans certains cas, d'évaluer la pertinence de la relation de linéarité qui doit relier le logarithme du risque instantané de décès aux différentes covariables.

Ainsi, si l'on pense qu'il existe un seuil de fonction rénale au-delà duquel le pronostic est brutalement aggravé (ce qui traduit par définition une non-linéarité), une représentation graphique des résidus en fonction de la variable « log_urée » pourrait avoir l'allure :

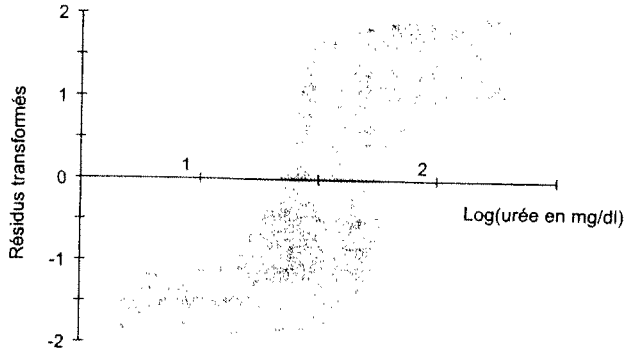

Fig. 3.2 — *Allure des résidus en cas d'écart flagrant à l'hypothèse de loglinéarité.*

Or la représentation graphique que nous obtenons est la suivante :

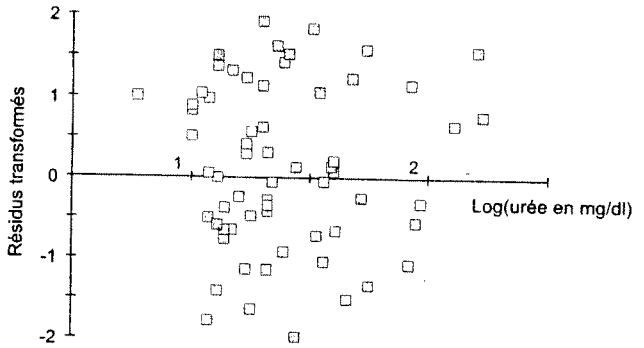

Fig. 3.3 — *Résidus exprimés en fonction de la covariable « log_urée ».*

Un tel écart à la linéarité n'est donc pas flagrant... Il faut toutefois noter qu'une telle approche est très peu sensible.

Recherche de sujets marginaux

En quelques mots : il s'agit ici de déterminer les sujets « influents », c'est-à-dire des sujets qui, une fois écartés du jeu de données, sont à l'origine des variations les plus importantes dans l'estimation des coefficients de régression.

En pratique : pour l'exemple que nous étudions, nous obtenons à l'aide de R le schéma suivant :

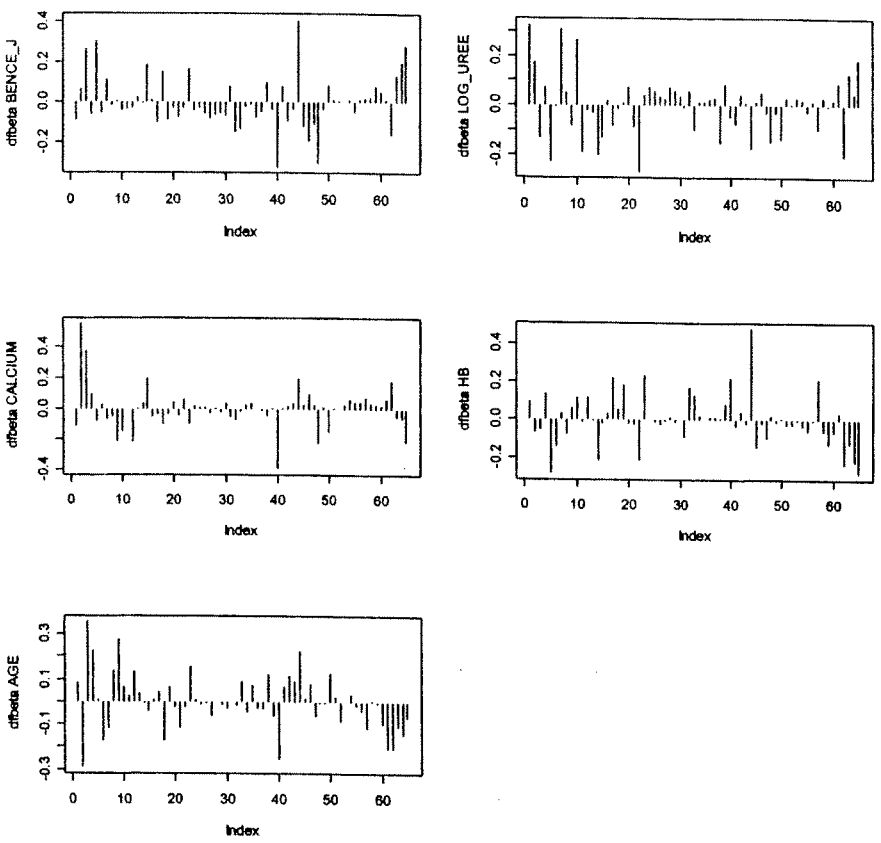

Fig. 3.4 — *Recherche de sujets influants pour un diagnostic de modèle de Cox.*

Le sujet 40 est influant pour les variables « Bence_J », « calcium » et « âge ». Le sujet 44 est influant pour les variables « Bence_J » et « Hb », etc. Il est alors possible d'extraire ces quelques sujets du jeu de données et d'estimer de nouveau le modèle. Si, qualitativement, les mêmes résultats sont trouvés, les conclusions en sortiront alors renforcées : le modèle est dit robuste.

4. Mesures répétées

En recherche biomédicale, répéter la même mesure au cours du temps sur un groupe de sujets est assez fréquent. Ainsi, pour comparer deux antidépresseurs, les patients vont-ils être évalués lors de la première prise du traitement, puis, par exemple, toutes les semaines pendant deux mois. De même, dans une étude de type cohorte épidémiologique, par définition, les sujets sont-ils évalués à plusieurs reprises.

Nous allons voir au cours des pages suivantes que le caractère temporel de ces mesures pose des difficultés insoupçonnées. S'il existe aujourd'hui un arsenal sophistiqué de techniques statistiques permettant d'analyser des données temporelles binaires, quantitatives, censurées ou pas, ces techniques sont en pratique d'un usage délicat et doivent être maniées avec prudence, voire seulement quand cela est vraiment indispensable.

Nous aborderons dans le détail le cas plus simple d'une variable à expliquer quantitative. Les notions d'analyse de variance multivariée (MANOVA), d'analyse de variance sur mesures répétées, de modèle mixte, de résidus à covariance structurée seront vues à ce propos. Nous terminerons par le cas d'une variable à expliquer binaire et par celui des données de survie.

La variable à expliquer est quantitative

Les données sont corrélées

En quelques mots : si, dans un modèle linéaire, la nécessité d'une indépendance des termes résiduels peut sembler relever plus du pinaillage théorique que de la réalité pratique, il n'en est rien dans les faits. Nous allons voir à partir d'un exemple que ce type de situation est banal et conduit à des conclusions erronées.

En pratique : un examen à base de potentiels évoqués corticaux est réalisé dans deux groupes de sujets : un groupe de vingt-cinq patients asymptomatiques séropositifs au VIH et un groupe de dix-neuf témoins. Chez chacun de ces sujets, les amplitudes (variable « amplit ») de quatre ondes électroencéphalographiques : N1, N2, P3a et P3b (notées onde = 1, 2, 3 et 4) sont mesurées à approximativement 100 ms, 200 ms, 280 ms et 320 ms après le déclenchement d'un stimulus auditif. Il s'agit donc bien d'une situation de mesures répétées chez des individus identiques.

La question est de savoir s'il existe, en moyenne, une différence globale d'amplitude des différentes ondes dans les groupes « séropositif » (19 patients) et « témoin » (25 patients).