Université de Montpellier
HMMA210 Projet

Étude du modèle de Cox : Vérification des hypothèses et quelques extensions

---

Étudiantes :

Alleau Julie	Master 1 Biostatistique
Lidoine Juliette	Master 1 Biostatistique
Serre-Combe Chloé	Master 1 Biostatistique

Encadrante :

Brunel-Piccinini Élodie

Année universitaire 2020-2021

Remerciements

Nous remercions Élodie Brunel-Piccinini, professeure à l'Université de Montpellier et encadrante de notre projet, pour son soutien et sa disponibilité dont elle a fait preuve tout au long de nos recherches pour ce projet.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Le modèle de Cox ou modèle à risque proportionnel a été introduit en 1972 par le statisticien britannique David Cox (13). Ce modèle permet d'analyser le temps écoulé depuis une date d'origine jusqu'à ce qu'un événement ne survienne tout en modélisant de façon flexible les éventuelles données censurées. Il s'agit d'une des méthodes les plus utilisées dans l'analyse de survie puisqu'il permet d'inclure des covariables quantitatives ou qualitatives afin de quantifier le risque associé à chacune d'entre elles.

Le modèle de Cox est un modèle semi-paramétrique qui permet de modéliser la fonction de risque instantané. Depuis 1972, dans la littérature, il est possible de trouver énormément d'informations et d'applications autour de ce modèle et ses possibles extensions. En effet, le modèle de Cox est devenu très populaire du fait de l'interprétation de ses coefficients en terme de risque relatif à l'exposition à un ou plusieurs facteurs. Néanmoins, une fois l'estimation réalisée, la vérification des hypothèses sur lesquelles le modèle de Cox repose, est trop souvent négligée. Cependant, il existe différentes façons de vérifier si ces hypothèses sont vérifiées ou non et il existe des alternatives pour utiliser le modèle de Cox lorsque celles-ci ne sont pas valides.

Après une brève présentation d'un jeu de données réelles `bfeed` (Klein et Moeschberger (1997) (15)), issues du package `KMsurv` du logiciel R qui nous serviront à illustrer nos résultats, nous proposerons un premier chapitre qui définit le modèle de Cox et nous rappellerons comment les estimateurs du maximum de vraisemblance sont obtenus. Nous verrons également comment choisir le "meilleur" modèle à l'aide d'une méthode et de critères de sélection de modèle. Ensuite, dans un deuxième chapitre, nous donnerons les hypothèses du modèle de Cox et comment vérifier leur validité. Enfin, dans un troisième et dernier chapitre, nous nous intéresserons à la manière dont des covariables dépendant du temps peuvent être intégrées au modèle de Cox.

PRÉSENTATION DES DONNÉES

Dans la suite de ce rapport, nous allons illustrer nos propos avec les données `bfeed` que l'on retrouve dans le package `KMSurv`. Ces données ont été recueillies lors d'une étude sur 927 femmes allaitantes aux États-Unis. Le but de cette étude est de mesurer la durée de l'allaitement, en semaines, chez ces femmes, en fonction de plusieurs covariables (l'événement d'intérêt est donc ici l'arrêt de l'allaitement). Ces durées sont répertoriées dans `duration` et les censures à droite (non observation de l'événement d'interêt) sont répertoriées dans la variable `delta`. Nous avons des observations dépendant des dix covariables suivantes :

Covariables qualitatives :

- `delta` : est l'indicateur de censure, égale à 1 si l'arrêt de l'allaitement est observé, et 0 sinon ;
- `smoke` : représente la consommation de tabac de la mère, égale à 1 si la mère est fumeuse et 0 sinon ;
- `alcohol` : représente la consommation d'alcool de la mère, égale à 1 si la mère en consomme et 0 sinon ;
- `pc3mth` : représente la nécessité de soins prénataux après le troisième mois de grossesse, égale à 1 si il y a eu recours à des soins, 0 sinon ;
- `race` : représente la couleur de peau de la mère, égale à 1 pour une couleur de peau blanche, 2 pour une couleur de peau noire et 3 pour une couleur de peau "autre".

Covariables quantitatives discrètes :

- `agemth` : représente l'âge de la mère à la naissance de l'enfant ;
- `ybirth` : représente l'année de naissance de l'enfant ;
- `yschool` : représente le nombre d'années d'études de la mère.

Nous allons nous concentrer sur les covariables `race`, `smoke`, `yschool` et `ybirth` qui apparaîtront lors de l'analyse statistique de la survie, comme les variables explicatives ayant le plus d'influence sur la durée de vie. Faisons alors une étude descriptive préliminaire de ces dernières.

Tout d'abord nous souhaitons savoir quel est le taux de censure de nos données. Pour cela nous pouvons le calculer grâce à la formule suivante :

$$1 - \frac{1}{n}\sum_{i=1}^{n} \delta_i.$$

Le taux de censure dans nos données est donc de 0.038, ce qui est relativement faible.

Voici deux tableaux résumant les effectifs des deux covariables qualitatives `smoke` et `race`.

fumeuse	**non-fumeuse**
270	657

TABLE 1 – Effectif des femmes pour chaque modalité de `smoke`

La proportion de femmes fumeuses est ici de 29% et on a donc 71% de femmes non fumeuses.

blanc	**noir**	**autres**
662	117	148

TABLE 2 – Effectif des femmes pour chaque modalité de `race`

La proportion de femmes de couleur de peau blanche dans cette étude est de 71% contre 13% de femmes de couleur de peau noire et 16% d'autres couleurs de peau.

Présentons maintenant plus en détail la covariable `yschool` à l'aide des tableaux suivants.

3	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
1	4	6	18	37	66	88	438	88	76	26	66	5	5	3

TABLE 3 – Effectif des femmes pour chaque modalité de `yschool`

Min.	**1er Qu.**	**Median**	**moy.**	**3ème Qu.**	**Max.**
3.00	12.00	12.00	12.21	13.00	19.00

TABLE 4 – Résumé statistique de la covariable `yschool`

Enfin, voici les détails statistiques de la covariable `ybirth` dans les deux tableaux suivants.

78	**79**	**80**	**81**	**82**	**83**	**84**	**85**	**86**
55	91	103	147	125	142	123	130	11

TABLE 5 – Effectif des femmes pour chaque modalité de `ybirth`

Min.	**1er Qu.**	**Median**	**moy.**	**3eme Qu.**	**Max.**
78.00	80.00	82.00	81.97	84.00	86.00

TABLE 6 – Résumé statistique de la covariable `ybirth`

Afin de mieux comprendre ces covariables regardons de plus ces quelques graphes.

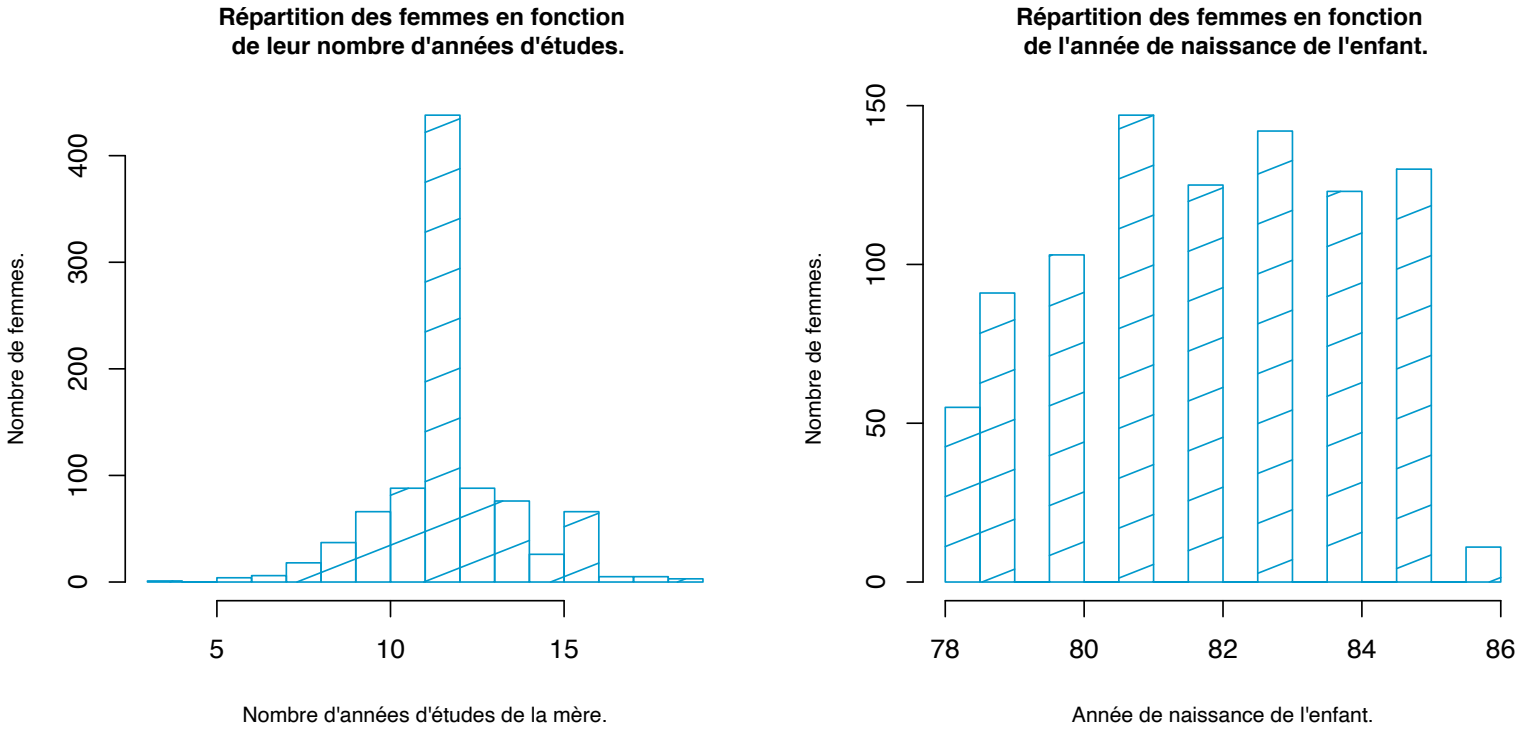

FIGURE 1 – Représentation des covariables `ybirth` et `yschool`.

Sur le 1^{er} graphique représentant la covariable `yschool`, on remarque un nombre beaucoup plus important de femmes qui ont fait 12 années d'études que dans les autres modalités. Cela s'explique car ces 12 années correspondent au nombre d'années total d'études aux États-Unis de l'école primaire au secondaire (ce qui est équivalent à la terminale en France). On a alors, pour 12 ans, 269 femmes sur 927 soit 29% des femmes de l'étude qui ont fait des études supérieures.

Sur le second graphique représentant la covariable `ybirth`, on remarque qu'il y a beaucoup moins de femmes donnant naissance en 1986 dans l'étude (seulement 11 sur 927) et que ce sont les femmes donnant naissance en 1981 qui ont davantage participé à l'étude (147 sur 927).

Regardons à présent, des diagrammes représentant les proportions de femmes pour les covariables `smoke` et `race`.

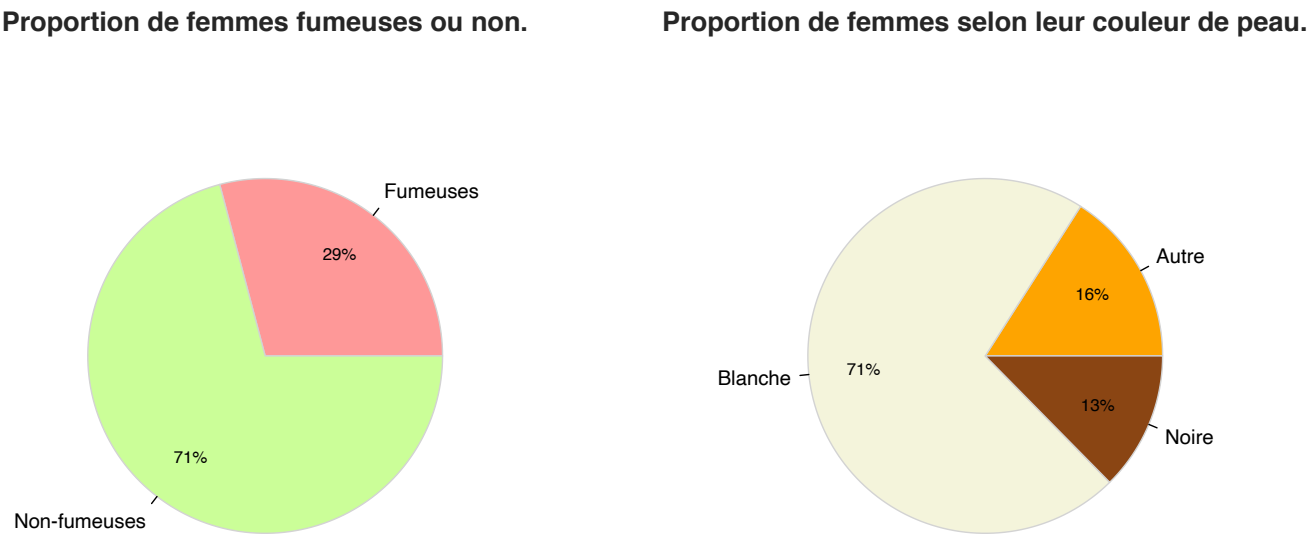

FIGURE 2 – Représentation des covariables `race` et `smoke`.

Sur le diagramme circulaire représentant la proportion de femmes fumeuses, on remarque qu'il y a un pourcentage plus important de femmes non fumeuses dans notre étude (71%).

Sur le deuxième diagrammme circulaire représentant la proportion de femmes selon leur couleur de peau, on remarque qu'il y a un pourcentage beaucoup plus élevé de femmes blanches que de femmes de toute autre couleur de peau dans cette étude.

Regardons maintenant la courbe de survie selon Kaplan-Meier afin de mieux analyser les données de la variable `duration`.

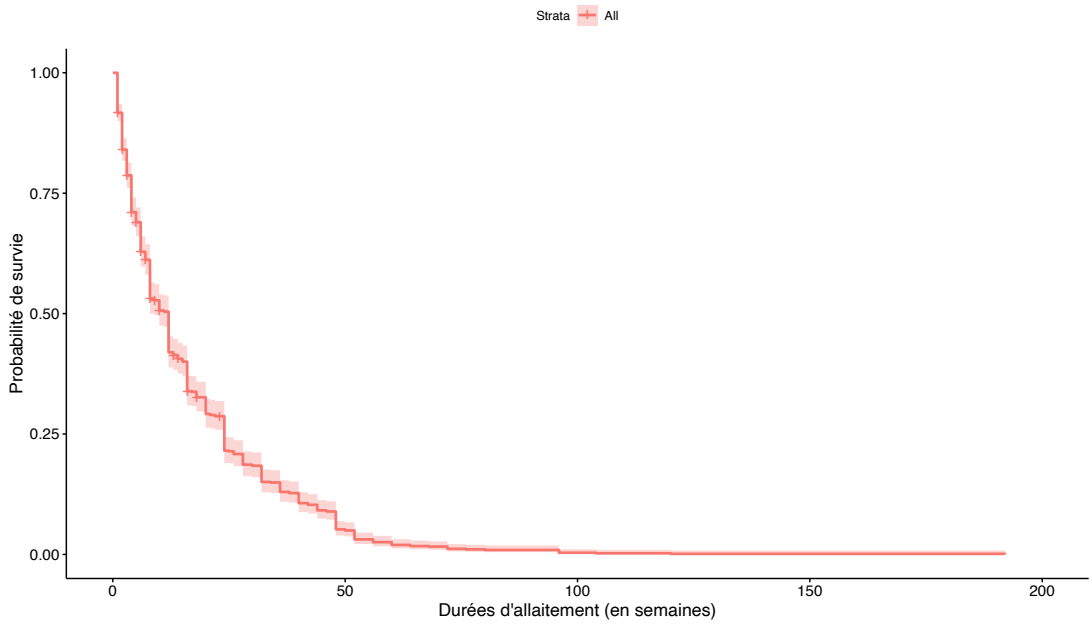

FIGURE 3 – Estimation de la courbe de survie selon Kaplan-Meier.

Sur ce dernier graphique on remarque qu'après 50 semaines, la probabilité de continuer d'allaiter est quasi nulle. On remarque également qu'il y a des censures au début de l'étude mais passé les 25 semaines, il n'y en a plus.

Voici un tableau présentant les aspects statistiques des durées d'allaitement avec censure :

Min.	**1er Qu.**	**Median**	**moy.**	**3eme Qu.**	**Max.**
1.00	4.00	10.00	16.18	24.00	192.00

TABLE 7 – Résumé statistique de la variable de la durée d'allaitement `duration` avec censures.

Ce tableau montre qu'il n'y a pas beaucoup de données au delà de 24 semaines. En effet, au moins 75% des valeurs sont inférieures ou égales au $3^{ème}$ quartile correspondant à 24 semaines. On remarque aussi que les durées d'allaitement sont très étalées dans le temps. Effectivement, la durée minimum d'allaitement est de 1 semaine contre 192 semaines soit environ 3 ans et demi pour la durée d'allaitement maximale.

Regardons maintenant un tableau présentant les aspects statistiques des durées d'allaitement sans censure :

Min.	**1er Qu.**	**Median**	**moy.**	**3eme Qu.**	**Max.**
1.00	4.00	10.00	16.51	24.00	192.00

TABLE 8 – Résumé statistique de la variable de la durée d'allaitement `duration` sans censure.

Ce tableau est très similaire à la `Table 7`, seule la moyenne est différente. En effet, ces données ne comptent que 35 censures soit une proportion de 0.038 ce qui est relativement faible.

CHAPITRE 1

MODÈLE DE COX

Le modèle de Cox est un modèle de régression semi-paramétrique en temps continu. Comme tout modèle de régression, ce modèle admet des variables explicatives, notées Z_i. Nous verrons par la suite que l'effet de ces variables explicatives peut ne pas être constant au cours du temps ce qui change la manière de procéder. Mais tout d'abord intéressons nous aux quelques définitions suivantes.

1.1 Définitions

La régression est une méthode statistique qui permet d'analyser la relation d'une variable par rapport à d'autres variables étant corrélées entre elles. Il existe différentes régressions telles que la régression linéaire, la régression non linéaire, la régression logistique ou encore la régression non paramétrique.

Définition 1.1.1 (Modèle de régression)**.**
Un modèle de régression paramétrique s'écrit sous la forme $Y = f_\beta(X) + \varepsilon$, (Nguyen, 2014, (11)) avec Y le vecteur des valeurs de la variable aléatoire réelle à expliquer, X la matrice contenant les valeurs des variables explicatives du modèle, f_β la fonction de régression paramétrée par un vecteur de coefficients β (celle-ci est affine pour une régression linéaire) et ε le vecteur des erreurs aléatoires.

Ce modèle est utilisé pour analyser la relation entre deux ou plusieurs variables et permet ensuite d'estimer une variable en se basant sur les autres.

Modèle semi-paramétrique

La régression non paramétrique correspond à une approche dans laquelle on ne fait aucune hypothèse sur les lois, ainsi f_β devient f. A l'inverse, dans la régression paramétrique, des hypothèses sur les lois sont effectuées.

Le modèle de régresson semi-paramétrique va alors être une forme intermédiaire entre le modèle de régression paramétrique et le modèle non paramétrique. Ce modèle est plus flexible qu'un modèle paramètrique avec moins d'hypothèses à vérifier.

Définition 1.1.2 (Fonction de risque instantané)**.**
La fonction de risque instantané, pour t fixé, est la probabilité de mourir dans un intervalle de temps $[t, t+a]$, sachant que l'on est encore en vie à l'instant t. Celle-ci est définie de la manière suivante :

$$h(t) = \lim_{a \to 0} \frac{\mathbb{P}(t \leq X < t+a | X \geq t)}{a} = \frac{f(t)}{S(t)} = -\ln(S(t))',$$

avec f la fonction de densité et S la fonction de survie définie par

$$S(t) = \mathbb{P}(X \geq t).$$

Définition 1.1.3 (Modèle de Cox)**.**
Le modèle de Cox est un modèle de régression semi-paramétrique permettant d'exprimer le risque instantané de décès en fonction de l'instant t et des covariables. Les covariables sont des variables explicatives quantitatives ou qualitatives, représentant des facteurs de risque, des facteurs pronostiques, ou des groupes auxquels appartiennent les sujets de l'étude.
Le modèle s'écrit :

$$h(t|Z) = h_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2, ... + \beta_p Z_p), \ \forall t \geq 0, \tag{1.1}$$

où h_0 est appelée fonction de risque de base, $\beta = (\beta_1, ..., \beta_p)^T$ est le vecteur des coefficients et $Z = (Z_1, ..., Z_p)^T$ est le vecteur des covariables.

Dans cette définition, ce sont les β_i qui sont les plus importants. En effet, l'interprétation se fait en fonction des $\exp(\beta_i)$ qui représentent le risque relatif (noté RR) associé à la covariable Z_i, toutes les autres covariables étant égales par ailleurs. Pour le risque associé à la covariable Z_i on a pour tout $l, k \in 1, ..., p$, avec $Z_{i,l}$ désignant la valeur de la covariable Z_i pour l'individu l,

$$\begin{aligned} RR &= \frac{h(t|Z_{i,l})}{h(t|Z_{i,k})} \\ &= \frac{h_0(t) \exp(\beta_1 Z_{1,l} + ... + \beta_i Z_{i,l} + ... + \beta_p Z_{p,l})}{h_0(t) \exp(\beta_1 Z_{1,k} + ... + \beta_i Z_{i,k} + ... + \beta_p Z_{p,k})}, \end{aligned}$$

en considérant toutes les autres covariables comme étant égales on a,

$$\begin{aligned} &= \frac{\cancel{h_0(t)} \exp(\cancel{\beta_1 Z_{1,l}} + ... + \beta_i Z_{i,l} + ... + \cancel{\beta_p Z_{p,l}})}{\cancel{h_0(t)} \exp(\cancel{\beta_1 Z_{1,k}} + ... + \beta_i Z_{i,k} + ... + \cancel{\beta_p Z_{p,k}})} \\ &= \frac{\exp(\beta_i Z_{i,l})}{\exp(\beta_i Z_{i,k})} \\ &= \exp(\beta_i (Z_{i,l} - Z_{i,k})). \end{aligned}$$

Pour des variables qualitatives ou binaires, la valeur $\exp(\beta_i)$ donne le risque relatif de décés (événement d'intérêt) à l'exposition du facteur Z_i d'un individu par rapport à un autre individu possédant la modalité de référence.

Application sur les données `bfeed` : Soit le modèle avec $Z_i =$ `smoke`, à deux modalités. La valeur 0 est la modalité de référence pour une variable binaire :

- Femmes fumeuses $Z_{i,l} = 1$;
- Femmes non fumeuses $Z_{i,k} = 0$.

Si nous prenons les femmes non fumeuses comme modalité de réference, alors

$$\frac{h(t|Z_{i,l})}{h(t|Z_{i,k})} = \exp(\beta_i(1-0)) = \exp(\beta_i).$$

Donc $\exp(\beta_i)$ correspond au risque d'arrêt de l'allaitement d'une femme fumeuse par rapport à une femme non fumeuse.

Pour des variables quantitatives, les valeurs $\exp(\beta_i)$ donnent le risque lié à l'augmentation d'une unité de la covariable.

Application sur les données `bfeed` : Soit le modèle avec $Z_i =$ `yschool` on a,

$$\begin{aligned}\frac{h(t|Z_{i,l})}{h(t|Z_{i,k})} &= \exp\left(\beta_i(Z_{i,l} - Z_{i,k})\right) \\ &= \left(\exp(\beta_i)\right)^{(Z_{i,l}-Z_{i,k})}.\end{aligned}$$

Donc si $Z_{i,l}$ et $Z_{i,k}$ diffèrent d'une unité, par exemple $Z_{i,l} = n+1$ années et $Z_{i,k} = n$ années, alors le risque lié à l'augmentation d'une unité de la covariable `yschool` est $(\exp(\beta_i))^{(Z_{i,l}-Z_{i,k})} = (\exp(\beta_i))^{n+1-n} = \exp(\beta_i)$.

Remarque :
Si $\beta_i < 1$, c'est une diminution du risque.
Si $\beta_i > 1$, c'est une augmentation du risque.

Regardons maintenant quelques caractéristiques de β.

- Si $\beta = 0$ alors $RR = 1$, cela signifie qu'il n'y a pas de relation entre la variable $h(t)$ et Z. Par exemple, pour nos données sur l'allaitement, si $\beta_1 = 0$ et $Z_1 =$ `smoke` alors il n'y aurait pas de lien entre la durée de l'allaitement et la covariable `smoke` (le fait de fumer n'influerait pas sur la durée de l'allaitement) ;
- Si $\beta < 0$ alors $RR < 1$, cela signifie que la covariable associée à moins de risque de "décés" que la variable de référence ;
- Si $\beta > 0$ alors $RR > 1$, à l'inverse, cela signifie que l'exposition à un tel facteur augmente le risque de "décés".

Nous avons dit précédemment que si $\beta = 0$ alors la covariable associée au β n'influe pas sur l'événement d'intérêt. Ainsi, il serait inutile de l'utiliser dans notre modèle. C'est pour cela que nous avons besoin de faire des tests sur ces coefficients β.

1.2 Vraisemblance

Le modèle de Cox est estimé en utilisant le principe du maximum de vraisemblance. Comme le terme $h_0(t)$ est considéré comme du "bruit" (ne dépend pas de β) , il ne sera pas estimé. On maximisera donc une log-vraisemblance partielle. Pour cela nous avons besoin de la vraisemblance qui est définie par :

$$L(\beta) = \prod_{j=1}^{D} \frac{\exp(\sum_{l=1}^{p} \beta_l Z_{(j),l})}{\sum_{i \in R_j} \exp(\sum_{l=1}^{p} \beta_l Z_{i,l})} = \prod_{j=1}^{D} \frac{\exp(Z_j^T \beta)}{\sum_{i \in R_j} \exp(Z_i^T \beta)}. \tag{1.2}$$

En passant au log nous obtenons donc :

$$\log(L(\beta)) = \sum_{j=1}^{D} \exp(Z_j^T \beta) - \log\left(\sum_{i \in R_j} \exp(Z_i^T \beta)\right).$$

avec :

- D le nombre de "décès" observés ;
- $Z_{(j),l}$ est la valeur de la covariable l de l'individu qui "décède" en $t_{(j)}$;
- R_j l'ensemble des individus exposés au risque de "décès" en $t_{(j)}$, avec $t_{(i)} \geq t_{(j)}$;
- $Z_{i,l}$ est la valeur de la covariable l de l'individu i.

Le modèle de Cox permet de traiter des données en temps continu. Néanmoins dans la pratique, il arrive souvent que plusieurs observations se produisent à la même date. Dans ce cas, il existe deux méthodes pour calculer la vraisemblance partielle qui sont :

1) **La méthode de Breslow** (1974, (12)) : La vraisemblance partielle devient alors :

$$L^*(\beta) = \prod_{j=1}^{D} \frac{\exp(\sum_{i \in D_j} (Z^{(i)})^T \beta)}{\left(\sum_{i \in R_j} \exp((Z^{(i)})^T \beta)\right)^{m_j}}, \tag{1.3}$$

avec :

- $Z^{(i)} = (Z_{1,i}, ..., Z_{p,i})^T$ le vecteur des p covariables du patient i ;
- D_j l'ensemble des indices des individus qui "décèdent" en $t_{(j)}$;
- m_j le nombre de "décès" observés en $t_{(j)}$, avec $m_j \geq 1$.

2) **La méthode d'Efron** (1977, (14)) : Il s'agit d'une approximation plus fine, la vraisemblance partielle devient alors :

$$L^*(\beta) = \prod_{j=1}^{D} \frac{\exp(\sum_{i \in D_j} (Z^{(i)})^T \beta)}{\prod_{l=1}^{m_j} \left(\sum_{i \in R_j} \exp((Z^{(i)})^T \beta) - \frac{l-1}{m_j} \sum_{i \in D_j} \exp((Z^{(i)})^T \beta)\right)}.$$

S'il n'y a pas d'ex-aequos toutes les méthodes sont équivalentes.
Comme la vraisemblance partielle ne dépend pas de la fonction de risque de base $h_0(t)$, on peut estimer β sans connaître cette fonction, ce qui est un atout car nous ne connaissons pas cette fontion en pratique.

On estime les coefficients β par maximisation de la vraisemblance partielle de Cox. Ainsi, on peut utiliser des propriétés asymptotiques et les utiliser pour estimer et tester ces coefficients.

Définition 1.2.1 (Estimation des paramètres)**.**
L'estimateur $\hat{\beta}$ du maximum de vraisemblance de β vérifie :

$$\frac{\partial \log(L(\hat{\beta}))}{\partial \beta} = 0.$$

L'information de Fisher de β est :

$$\mathcal{I}_\beta = -\mathbb{E}\left[\frac{\partial^2 \log(L(\beta))}{\partial \beta_i \partial \beta_j}\right]_{i,j=1,\dots p},$$

où $\left[\frac{\partial^2 \log(L(\beta))}{\partial \beta_i \partial \beta_j}\right]_{i,j=1,\dots p}$ correspond au coefficient i, j de la matrice hessienne.
La matrice de variance-covariance de $\hat{\beta}$ est :

$$V(\hat{\beta}) = \mathcal{I}_\beta^{-1}.$$

Comme $\hat{\beta}$ est un estimateur obtenu par maximum de vraisemblance (partielle), on a la propriété suivante :

Propriété 1 (Normalité asymptotique)**.**
On a

- $\hat{\beta} \xrightarrow{p.s} \beta$;
- $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0_p, \mathcal{I}_\beta^{-1})$.

Maintenant que nous savons comment estimer les paramètres β d'un modèle, nous pouvons nous demander quelles sont les covariables utiles à étudier. En effet, dans un modèle à plusieurs covariables, celles-ci ne sont pas toutes nécessaires pour étudier des données ou encore faire de la prédiction. Dans la partie suivante nous allons donc montrer plusieurs méthodes pour choisir un modèle pertinent afin de se concentrer sur l'estimation des paramètres de ce modèle simplifié ou non.

1.3 Critères de sélection

Nous nous retrouvons ainsi avec un modèle multivarié contenant des covariables potentiellement corrélées entre elles. Nous allons alors chercher à sélectionner les variables les plus pertinentes. Pour cela, à l'aide des estimateurs du maximum de vraisemblance, une façon classique de sélectionner un meilleur modèle est de pénaliser la vraisemblance. Le modèle sera alors à la fois plus performant avec les résidus les plus petits possibles et plus économique, c'est-à-dire avec le moins de variables explicatives possible.

Pour ce faire, il existe différents critères et méthodes de sélection. Nous allons vous présenter le critère AIC, le critère BIC et la méthode descendante.

1.3.1 Le critère d'information d'Akaike (AIC)

Définition 1.3.1 (Le critère AIC)**.**
Le critère d'information d'Akaike (AIC) s'applique sur un ensemble de modèles candidats et repose sur une méthode du maximum de vraisemblance (Bertrand et Maumy, 2008, (8)) en introduisant des mesures de pénalités dans la log-vraisemblance.
Il est défini ainsi :

$$AIC(M_p) = -2\log(\tilde{L}_p) + 2p,$$

où M_p est un modèle à p paramètres et $\tilde{L}_p$ est la vraisemblance maximisée du modèle M_p.

Propriété 2 (Sélection du modèle)**.**
Lors de l'application du critère AIC sur un ensemble de modèles candidats, le meilleur modèle (celui qui sera choisi) sera celui possédant l'AIC le plus faible. Il s'agit donc du modèle M_{AIC} tel que :

$$M_{AIC} = \arg\min_{M_p \in \mathcal{M}} AIC(M_p),$$

où $\mathcal{M}$ est l'ensemble de tous les modèles possibles.

L'AIC réalise le meilleur compromis biais-variance (Saporta, 2012, (9)) en prenant en compte à la fois la qualité de l'ajustement et la complexité du modèle. Il va pénaliser les modèles ayant un grand nombre de paramètres afin de limiter d'éventuels problèmes de sur-ajustement. En effet, un sur-ajustement pourrait entraîner une considération de paramètres qui n'ont, en réalité, pas d'influence sur le modèle général.

Algorithme de l'AIC

Faire le choix du meilleur modèle en utilisant le critère AIC consiste à partir du modèle complet et d'éliminer les variables une à une lorsque l'élimination de celles-ci permet de diminuer l'AIC. Ainsi, on va considérer des familles de modèles emboités et pour chaque niveau on conserve le modèle qui minimise l'AIC.

Application avec les données `bfeed` : Nous avons donc utilisé ce critère sur nos données `bfeed` et nous obtenons un tableau contenant les covariables retenues par cette méthode. Le voici :

Modèles	paramètres	AIC
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `agemth`, `pc3mth`, `alcohol`	8	10351.97
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `agemth`, `alcohol`	7	10350.35
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `alcohol`	6	10349.11
`race`, `smoke`, `ybirth`, `yschool`, `poverty`	5	**10348.72**
`race`, `smoke`, `ybirth`, `yschool`	4	10352
`race`, `smoke`, `ybirth`, `poverty`	4	10356
`smoke`, `ybirth`, `yschool`, `poverty`	4	10357
`race`, `ybirth`, `yschool`, `poverty`	4	10357
`race`, `smoke`, `yschool`, `poverty`	4	10362

TABLE 1.1 – Critère AIC sur les données `bfeed`

Nous remarquons alors qu'avec la méthode du critère AIC nous retenons un modèle avec les 5 covariables `poverty`, `yschool`, `race`, `smoke` et `ybirth` qui est celui qui minimise l'AIC à 10348.72.

1.3.2 Le Bayesian Information Criterion (BIC)

Définition 1.3.2 (Le critère BIC)**.**
Le Bayesian Information Criterion (BIC) est défini ainsi :

$$BIC = -2\log(\tilde{L}_p) + p\log(n),$$

où $\tilde{L}_p$ est la vraisemblance maximisée, p est le nombre de paramètres dans le modèle et n le nombre d'observations.

Propriété 3 (Sélection du modèle)**.**
Lors de l'application du critère BIC sur un modèle, le meilleur modèle sera celui possédant le BIC le plus faible. Il s'agit donc du modèle M_{BIC} tel que :

$$M_{BIC} = \arg\min_{M \in \mathcal{M}} BIC(M).$$

À la différence du critère AIC, la pénalité dépend de la taille de l'échantillon et pas seulement du nombre de paramètres. De plus, le critère BIC va permettre de retenir les variables les plus statistiquement significatives du modèle (Bertrand et Maumy, 2008, (8)).

Algorithme du BIC

Faire le choix du meilleur modèle en utilisant le critère BIC consiste, à partir du modèle complet, à éliminer les variables une à une lorsque l'élimination de celles-ci permet de diminuer le BIC. Ainsi, on va considérer des familles de modèles emboités et pour chaque niveau on conserve le modèle qui minimise le BIC.

Application avec les données `bfeed` : De la même manière que pour le critère AIC, nous avons utilisé le critère BIC sur nos données `bfeed` et nous obtenons le tableau suivant :

Modèles	paramètres	BIC
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `agemth`, `pc3mth`, `alcohol`	8	10390.31
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `agemth`, `alcohol`	7	10383.91
`race`, `smoke`, `ybirth`, `yschool`, `poverty`, `alcohol`	6	10377.87
`race`, `smoke`, `ybirth`, `yschool`, `poverty`	5	10372.69
`race`, `smoke`, `ybirth`, `yschool`	4	**10370.7**
`race`, `smoke`, `ybirth`	3	10371
`smoke`, `ybirth`, `yschool`	3	10373
`race ybirth`, `yschool`	3	10373
`race`, `smoke`, `yschool`	3	10380

TABLE 1.2 – Critère BIC sur les données `bfeed`

Nous remarquons ici qu'avec la méthode du critère BIC, le modèle qui minimise ce critère avec un BIC de 10370.7, est le modèle avec les 4 covariables : `yschool`, `race`, `smoke` et `ybirth`. Nous voyons donc que ce critère est plus strict que le critère AIC puisque la covariable `poverty` n'a pas été retenue ici contrairement à la sélection précédente.

1.3.3 Méthode pas à pas descendante

Algorithme de la méthode descendante :

La méthode descendante consiste, à partir du modèle complet, à éliminer les variables une à une lorsque celles-ci sont les moins significatives à chaque étape. Dans le cadre du modèle de Cox, nous allons donc regarder les p-values du test de Wald. Nous obtenons alors un modèle final avec toutes ses covariables significatives selon un seuil choisi (par exemple 5%, 1% ou 0.1%).

Tout d'abord, définissons le test de Wald. Pour chaque coefficient, on effectue ce test qui vérifie la significativité du coefficient tel que :

$$\mathcal{H}'_0 : \beta_i = 0 \quad \text{vs} \quad \mathcal{H}'_1 : \beta_i \neq 0.$$

Grâce au théorème central limite, on a la statistique de test de Wald z telle que

$$z = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{V}(\hat{\beta}_i)}} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1),$$

avec $\hat{V}(\hat{\beta}_i)$ l'estimation de la variance de $\hat{\beta}$.
Les variances estimées sont les coefficients diagonaux de l'inverse de l'estimation de l'information de Fisher donc $\hat{V}(\hat{\beta}_i) = \left[\hat{\mathcal{I}}_\beta\right]_{i,i}$. L'estimation de l'information de Fisher, elle, s'obtient par l'algorithme de minimisation de la vraisemblance en faisant l'estimation de la matrice hessienne.
Donc sous $\mathcal{H}'_0$, la statistique de test de Wald est

$$z = \frac{\hat{\beta}_i}{\sqrt{\hat{V}(\hat{\beta}_i)}} \sim \mathcal{N}(0,1),$$

La p-value du test de Wald, avec z_{obs} la valeur observée dans les données, est

$$\mathbb{P}(z^2 > z_{obs}) \text{ avec } z^2 \sim \chi^2_1.$$

De plus, un intervalle de confiance de niveau $1-\alpha$ pour chaque risque $\exp(\beta_i)$ est :

$$IC_{1-\alpha}(\exp(\beta_i)) = \left[\exp\left(\hat{\beta}_i - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{V}(\hat{\beta}_i)}\right); \exp\left(\hat{\beta}_i + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{V}(\hat{\beta}_i)}\right)\right],$$

où $z_{(1-\frac{\alpha}{2})}$ est le quantile d'ordre $1-\alpha$ de la loi normale.

Application avec les données `bfeed` : Essayons cette fois la méthode pas à pas descendante sur nos données `bfeed`. Nous obtenons les tableaux des covariables sélectionnées suivant :

Covariables	p-value
`race`	0.00113
`smoke`	0.00103
`ybirth`	7.78e-05
`yschool`	0.00185
`poverty`	0.03085

TABLE 1.3 – Méthode descendante avec poverty

En s'arrêtant lorsque toutes les p-values sont inférieures à 0.05, la méthode descendante nous donne un modèle à 5 covariables : `yschool`, `race`, `smoke`, `ybirth` et `poverty`. Ce qui revient alors à ce que nous avions obtenu à l'aide du critère AIC.

Néanmoins, si on décidait d'être plus stricte dans le choix des covariables et que nous décidions de garder seulement les covariables avec des p-values inférieures à 1% alors le modèle obtenu est celui avec les 4 covariables `yschool`, `race`, `smoke` et `ybirth`. Ce qui revient à ce que nous avions obtenu à l'aide du critère BIC.

Dans la suite, nous choisirons alors de garder le modèle "le plus simple" à 4 covariables obtenu par la méthode descendante et le critère BIC, afin de simplifier notre modèle et parce que nous avons peu d'informations complémentaires sur la covariable `poverty`.

1.4 Test d'hypothèses

Afin de tester la significativité des coefficients, des tests d'hypothèses sont effectués. Un test d'hypothèses permet de fournir une règle de décision (rejeter ou ne pas rejeter une hypothèse faite sur un ou des paramètres du modèle) à partir de l'étude d'un modèle. Pour analyser le modèle de Cox, un test d'hypothèses global sur tous les paramètres sera alors effectué ainsi qu'un test pour chacun des paramètres.

On pose le test d'hypothèses global suivant :

$$\mathcal{H}_0 : \beta_1 = \beta_2 = ... = \beta_p = 0,$$
(hypothèse nulle)
vs
$$\mathcal{H}_1 : \exists\ i \in \{1, ..., p\}, \beta_i \neq 0.$$
(hypothèse alternative)

Trois tests peuvent être effectués pour tester la nullité simultanée des coefficients :

- Test du rapport de vraisemblance ;
- Test de Wald ;
- Test du Score (ou Rao).

Ils sont basés sur 3 statistiques de tests différentes qui suivent asymptotiquement, sous $\mathcal{H}_0$, des lois du χ^2 à p degrés de liberté. La partie 2.1.5 sur les tests statistiques de Choukroun (2008, (10)) nous permet de définir plus en détail ces statistiques de test.

Notons β_0 l'estimateur du maximum de vraisemblance si l'on impose $\mathcal{H}_0$ (dans notre cas, $\beta_0 = 0_p$ où 0_p est le vecteur composé de p zéros) et notons $\hat{\beta}$ l'estimateur du maximum de vraisemblance non contraint.

Statistique de test de Wald :

Le test de Wald est basé sur la propriété 1 de la partie 1.2.
La statistique de test de Wald est

$$W_n = n(\hat{\beta} - \beta)^T \mathcal{I}_\beta (\hat{\beta} - \beta) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p.$$

Sous $\mathcal{H}_0$:

$$W_n = n(\hat{\beta} - \beta_0)^T \mathcal{I}_\beta (\hat{\beta} - \beta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p.$$

La p-value de ce test de Wald est :

$$\mathbb{P}\left[W_n > x^2_{p,1-\alpha}\right] = \alpha,$$

avec :

- $x^2_{p,1-\alpha}$ le quantile d'ordre $1-\alpha$ à p degrés de liberté de la loi χ^2 ;
- $\mathcal{I}_\beta$ l'information de Fisher définie à la définition **1.2.1** ;
- n le nombre d'individus.

Ainsi, on rejette $\mathcal{H}_0$ si $W_n > x^2_{p,1-\alpha}$ avec un risque α de se tromper, c'est-à-dire un risque de rejeter à tort $\mathcal{H}_0$.

Statistique de test du Score :

Notons U le vecteur des scores défini par :

$$\frac{\partial \log L(Z_i;\beta)}{\partial \beta},$$

où L est la vraisemblance définie dans la partie (1.2).
La statistique de test est :

$$R_n = \frac{1}{n} U(\beta_0)^T \mathcal{I}^{-1}_{\beta_0} U(\beta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p \quad (\text{sous } \mathcal{H}_0).$$

On rejette $\mathcal{H}_0$ si :

$$R_n > x^2_{p,1-\alpha}.$$

Statistique de test du rapport de vraisemblance :

On définit ainsi la statistique de test du rapport de vraisemblance :

$$\lambda_n = 2 \log \left(\frac{L(\hat{\beta})}{L(\beta_0)} \right),$$

L étant toujours la vraisemblance définie précédemment.
Si $\mathcal{H}_0$ est vraie on a :

$$\lambda_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p.$$

L'hypothèse globale $\mathcal{H}_0$ correspond au modèle sans variable explicative. On cherche à vérifier si le modèle avec variables explicatives est significativement plus performant pour expliciter des facteurs de risque de décés, que le modèle sans variable explicative.

Interprétation avec la sortie R

Maintenant que nous savons quels sont les tests d'hypothèses, nous allons voir comment faire pour rejeter ou non l'hypothèse $\mathcal{H}_0$. Nous allons étudier la sortie R du modèle de Cox exécuté sur les covariables séléctionnées précédemment.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
`factor(race)2`	0.16182	1.17565	0.10347	1.564	0.11783
`factor(race)3`	0.27842	1.32104	0.09710	2.867	0.00414 **
`factor(smoke)1`	0.24539	1.27812	0.07838	3.131	0.00174 **
`ybirth`	0.07191	1.07456	0.01792	4.013	6e-05 ***
`yschool`	-0.05016	0.95107	0.01915	-2.620	0.00880 **
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

	exp(coef)	exp(-coef)	inf.95	sup.95
`factor(race)2`	1.1756	0.8506	0.9599	1.4400
`factor(race)3`	1.3210	0.7570	1.0921	1.5980
`factor(smoke)1`	1.2781	0.7824	1.0961	1.4904
`ybirth`	1.0746	0.9306	1.0375	1.1130
`yschool`	0.9511	1.0514	0.9160	0.9874

	valeur	degré liberté	p-value
Test rapport vraisemblance	38.74	5	3e-07
Test de Wald	39.03	5	2e-07
Test du Score (logrank)	39.19	5	2e-07
Concordance[1]= 0.575 (se = 0.012)			
n= 927, nombre d'évènement d'intérêt= 892			

TABLE 1.4 – Sortie R de Coxph

Commençons d'abord par expliquer les termes de la sortie.
Nous avons :

- `coef` : la valeur prise par l'estimateur $\hat{\beta}_j$;
- `exp(coef)` : l'estimation du risque d'arrêt de l'allaitement ;
- `se(coef)` : l'écart-type estimé $\sqrt{\hat{V}(\hat{\beta})}$
- `z` : statistique de Wald du test d'hypothèses $\mathcal{H}'_0 : \beta_j = 0$ vs $\mathcal{H}'_1 : \beta_j \neq 0$;
- `Pr(>|z|)` : la p-value du test de Wald.

1. La concordance est un critère qui permet de juger le pouvoir *predictif* d'un modèle, c'est-à-dire qu'on regarde toutes les paires d'individus et on les classe en paires **concordantes** ou **discordantes**. Pour deux individus i et j si le modèle prédit que i a un risque de décès plus important que j et si $t_i < t_j$ alors la paire (i,j) est concordante. La formule est :

$$C = \frac{n_c + 0.5 n_t}{n_c + n_d + n_t}$$

avec n_c le nombre de concordances paires, n_d le nombre de discordances paires, et n_t le nombre de paires liées. (Breheny, 2019, (16))

Par exemple pour la covariable `ybirth` la probabilité de rejeter $\mathcal{H}'_0$ à tort est de $6e^{-05}$, ce qui signifie que l'on rejette $\mathcal{H}'_0$ très fortement puisque la probabilité de se tromper en rejetant est très faible. Ainsi, on peut penser que la covariable `ybirth` joue un rôle significatif dans le modèle et donc nous pouvons garder cette covariable dans le modèle.

Comme $\hat{\beta} > 0$, une année de naissance plus tardive augmente le risque d'arrêt de l'allaitement. Nous avons également $\exp(\hat{\beta}) = 1.074$ ce qui donne le risque à une unité x d'écart (ici x est une année) pour toute autre covariable égale par ailleurs. En effet, on a :

$$\exp(\hat{\beta}) = \exp(\hat{\beta}((x+1) - x)) = \frac{\exp(\hat{\beta}(x+1))}{\exp(\hat{\beta}x)}.$$

Donc si on considère une naissance en 1978 par rapport à une naissance en 1985 on a un risque d'arrêt d'allaitement de $\exp(\hat{\beta})^7 = (1.074)^7 = 1.65$. De ce fait, pour un écart d'année de naissance de 7 ans, on a environ $(1.65 - 1) = 65\%$ d'augmentation du risque d'arrêt de l'allaitement.

Pour finir, les p-values du test de la log-vraisemblance, du score ou encore de Wald sont inférieures à 1% donc on peut rejeter l'hypothèse globale $\mathcal{H}_0$. Ainsi, au moins un des coefficients est significativement non nul.

Nous avons alors vu comment se définit le modèle de Cox, de quelle manière sélectionner un modèle et enfin quels sont les tests d'hypothèses associés au modèle de Cox. Par ailleurs, pour utiliser ce modèle il faut vérifier deux hypothèses cruciales dont nous parlerons dans le chapitre suivant : l'hypothèse des hasards proportionnels et de la log-linéarité.

CHAPITRE 2

HYPOTHÈSES ET VÉRIFICATIONS

2.1 Hypothèses

Hypothèse des hasards proportionnels

Dans le modèle de Cox, le rapport des risques instantanés entre deux individus est constant *i.e.* ne dépend pas du temps. Le rapport mesurant ce risque relatif à un instant t donné pour un individu i par rapport à un individu j est donné par :

$$\frac{h(t|Z^{(i)})}{h(t|Z^{(j)})} = \frac{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_{i,k})}{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_{j,k})} = \exp((Z^{(i)} - Z^{(j)})^T\beta), \tag{2.1}$$

avec :

- $h(t|Z^{(i)})$ correspond au risque instantané pour l'individu i ;
- $\beta = (\beta_1, ..., \beta_p)^T$;
- $Z^{(i)} = (Z_{1,i}, ..., Z_{p,i})^T$ où $Z_{k,i}$ est la valeur de la covariable Z_k pour l'individu i ;
- $Z^{(j)} = (Z_{1,j}, ..., Z_{p,j})^T$ où $Z_{k,j}$ est la valeur de la covariable Z_k pour l'individu j ;
- $h_0(t)$ un risque de base qui est indépendant des variables explicatives du modèle.

Ainsi, ce rapport ne dépend pas du temps, il s'agit d'une constante dans le temps égale à $\exp((Z^{(i)} - Z^{(j)})^T\beta)$. Le modèle de Cox est alors basé sur un modèle à risques proportionnels. Néanmoins, en pratique, cette hypothèse n'est pas toujours vérifiée.

Hypothèse de log-linéarité

Dans le modèle de Cox, la deuxième hypothèse cruciale est l'hypothèse de log-linéarité qui suppose que l'on ait :

$$\log(h(t|Z)) = \log(h_0(t)) + \sum_{k=1}^{p}\beta_k Z_k. \tag{2.2}$$

Cela signifie que $\log(h(t|Z))$ est linéaire en fonction des Z_k. En pratique, cette hypothèse n'est pas toujours vérifiée.

Le modèle de Cox étudié jusqu'ici suppose que les covariables sont indépendantes du temps. Or, en pratique, ce n'est pas toujours le cas.

Par exemple, un état de santé, un lieu d'habitation ou encore une prise ou non d'un certain traitement sont des covariables susceptibles d'évoluer au cours du temps. On peut alors se demander, dans le cadre des données `bfeed`, si le fait de fumer ou le nombre d'années d'études sont des covariables qui peuvent être dépendantes du temps. En effet, une femme peut arrêter de fumer, stopper ou reprendre ses études au cours de l'étude. Il est donc important de vérifier la validité de ces hypothèses pour pouvoir appliquer le modèle de Cox.

2.2 Vérification de l'hypothèse des hasards proportionnels

Nous allons étudier comment vérifier l'hypothèse des hasards proportionnels. Celle-ci peut être vérifiée de plusieurs manières :

- Par des procédures graphiques ;
- Par des procédures de tests et graphiques sur les résidus de Schoenfeld, définis au paragraphe 2.2.2.

2.2.1 Représentation graphique

L'hypothèse des hasards proportionnels peut être vérifiée graphiquement.
En effet, on a :

$$\begin{aligned}
\log(-\log(S(t|Z^{(i)}))) &= \log(H(t|Z^{(i)})) \\
&= \log(\int_t^0 h(u|Z^{(i)})du) \\
&= \log(\int_t^0 h_0(u)\exp((Z^{(i)})^T\beta_i)du) \\
&= \log(\exp((Z^{(i)})^T\beta_i)\int_t^0 h_0(u)du) \\
&= \log(\exp((Z^{(i)})^T\beta_i)H_0(t)) \\
&= \log(H_0(t)) + (Z^{(i)})^T\beta_i.
\end{aligned}$$

où $H_0(t)$ est la fonction de risque cumulée de base.

Ceci signifie que, lorsque cette hypothèse est vérifiée, les courbes de survie $S(t)$ pour chaque covariable Z_i sont approximativement parallèles sur une échelle log-log.

Application avec les données `bfeed` : Comparons les courbes de survie selon les différentes modalités d'une covariable, ici `smoke`.

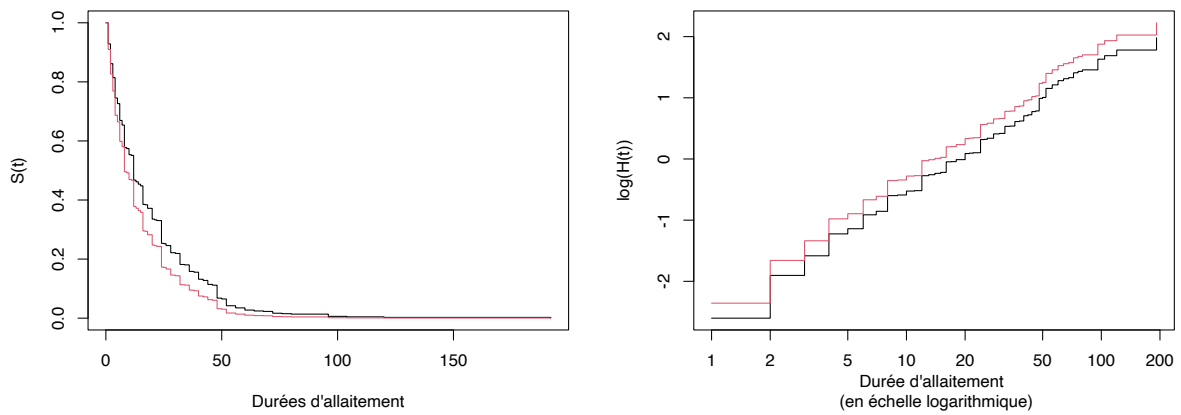

FIGURE 2.1 – Représentation des courbes de survie des femmes en fonction de leur consommation de tabac en échelle logarithmique ou non. En rouge, la courbe de survie des fumeuses, en noir, celle des non-fumeuses.

On remarque ici, que les courbes de survie pour les fumeuses (courbe rouge) et non-fumeuses (courbe noire) sont globalement parallèles en échelle logarithmique. On peut alors dire que l'hypothèse de proportionnalité est vérifiée graphiquement.

Cette méthode graphique est utilisable pour des covariables qualitatives ou quantitatives discrètes. Pour des variables continues, on pourra éventuellement les partitionner en groupes, ainsi on pourra les considérer comme des variables qualitatives ou quantitatives discrètes.

2.2.2 Résidus de Schoenfeld

Il est également possible de vérifier l'hypothèse de proportionnalité à l'aide des résidus de Schoenfeld, pour les covariables continues. En nous appuyant sur le travail de Berchtold (3), nous définissons les résidus de Schoenfeld de la manière suivante.

Définition 2.2.1 (Résidus de Schoenfeld)**.**
Le résidu de Schoenfeld, noté s_i, correspond à la différence entre la valeur observée de la covariable Z à l'instant t_i pour l'individu i et la valeur moyenne de la covariable Z pour tous les individus à risque à ce moment t_i. Ils sont donc calculés seulement pour les individus non-censurés.
Si les résidus de Schoenfeld sont distribués de la même manière dans le temps, l'hypothèse des hasards proportionnels est vérifiée. Sous l'hypothèse des risques proportionnels, la tendance des résidus de Schoenfeld est alors constante.

On utilise ici les résidus de Schoenfeld standardisés ce qui signifie qu'ils ont été divisés par leur variance.

Test de corrélation des résidus de Schoenfeld

Pour commencer, regardons le test de corrélation de ces résidus qui vérifie si une corrélation existe entre le résidu s_i et la durée t_i. Ce test du χ^2 pour la $j^{ème}$ covariable consiste à tester :

$$\mathcal{H}'_0 : \beta_j(t) = \beta_j \quad \text{contre} \quad \mathcal{H}'_1 : \beta_j(t) \neq \beta_j.$$

Cela signifie que pour chaque covariable Z on effectue la régression des résidus sur le temps et on teste la nullité de la pente via un test du χ^2. En effet, tester la nullité de la pente revient à tester le nullité du coefficient de corrélation.

Le test du χ^2 global consiste lui à tester :

$$\mathcal{H}_0 : \beta(t) = \beta \quad \text{contre} \quad \mathcal{H}_1 : \beta(t) \neq \beta.$$

Si l'on rejette l'hypothèse $\mathcal{H}_0$ alors on peut penser que certaines covariables ont un effet dépendant du temps.

Application : Effectuons ce test sur nos données.

	χ^2	df	p-value
`factor(race)`	4.462	2	0.1074
`factor(smoke)`	0.822	1	0.3646
`ybirth`	1.281	1	0.2577
`yschool`	8.436	1	0.0037
GLOBAL	12.237	5	0.0317

TABLE 2.1 – Résultat du test de corrélation des résidus.

Ce test montre qu'il existe un effet significativement dépendant du temps puisque la p-value pour la covariable `yschool` est inférieure à 0.05. On rejette alors l'hypothèse de proportionnalité au risque de 5%. Nous verrons dans le chapitre 3 comment étendre le modèle de Cox avec une covariable ne vérifiant pas cette hypothèse.

Représentation graphique des résidus de Schoenfeld

On peut ensuite réaliser la répresentation graphique de ces résidus en fonction du temps selon la covariable Z choisie. On retrouve, sur ce graphique, la tendance lissée des résidus, par splines (définis en 2.3.2), ainsi qu'un intervalle de confiance à 95%.
Afin de pouvoir interpréter ce graphique, il est intéressant d'ajouter l'effet estimé par le modèle de Cox. On dira alors que si la tendance lissée est approximativement la droite horizontale représentant l'effet estimé du modèle de Cox alors le risque est considéré comme constant et donc l'hypothèse des hasards proportionnels est validée.

Regardons alors la tendance lissée des résidus Schoenfeld avec nos données.

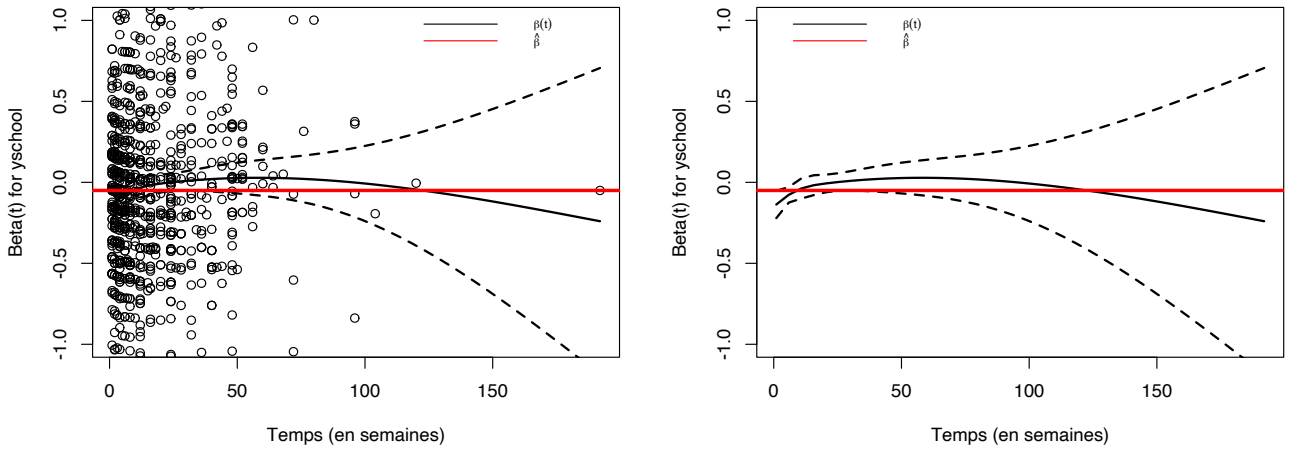

FIGURE 2.2 – Représentation des résidus de Schoenfeld de la covariable `yschool`, de leur tendance lissée ainsi que d'un intervalle de confiance à 95%.

La courbe noire en trait plein est la courbe de tendance lissée des résidus. La courbe en pointillés est l'intervalle de confiance à 95% de cette tendance lissée. Et la courbe rouge est l'effet estimé du modèle de Cox.

Pour la covariable `yschool`, la courbe noire est similaire à la courbe rouge on peut alors penser que le risque est constant et que l'hypothèse est validée. Ceci est en désaccord avec le test de corrélation réalisé précédemment où l'on avait rejeté $\mathcal{H}'_0$ et donc considéré que cette covariable était dépendante du temps.
Or ici, on regarde si $\beta(t) = \hat{\beta}$ et non si $\beta(t) = \beta$ comme nous l'avions fait lors du test de corrélation des résidus.
De plus, on remarque que l'intervalle de confiance de la tendance lissée devient large au fur et à mesure des semaines. Or, sur le graphe de gauche où les résidus sont visibles, on voit qu'une seule femme a allaité jusqu'à la semaine 192 et donc qu'il n'y a qu'un résidu à cet instant. Cet individu a donc un impact assez important sur la précision de cet intervalle de confiance.

Regardons maintenant la tendance lissée des résidus Schoenfeld de la covariable `smoke`.

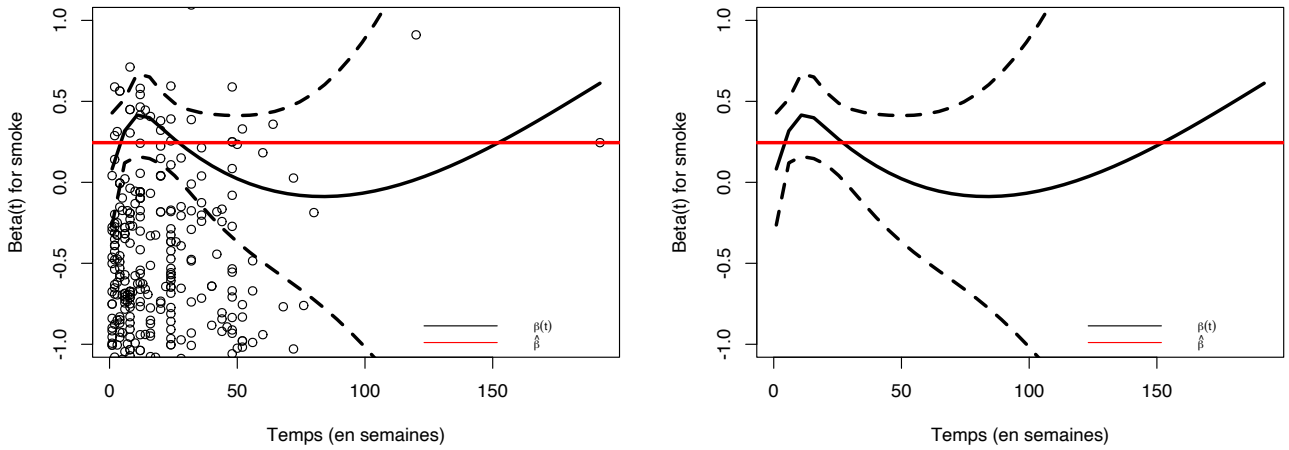

FIGURE 2.3 – Représentation des résidus de Schoenfeld de la covariable `smoke`, de leur tendance lissée ainsi que d'un intervalle de confiance à 95%.

Contrairement à la covariable `yschool`, la courbe noire en trait plein n'est pas vraiment similaire à la rouge pour la covariable `smoke`. On dira alors que le risque n'est pas constant dans le temps et que l'hypothèse de proportionalité n'est pas validée.
Ici également, puisque l'on sait que les femmes ayant allaité plus de 80 semaines sont peu nombreuses (8 exactement), l'intervalle de confiance de la tendance lissée des résidus n'est pas très précis. Nous verrons donc, dans le chapitre 3, comment prendre en compte cette covariable dans notre modèle de Cox malgré sa dépendance au temps.

En résumé, ces différentes méthodes de vérification de l'hypothèse de proportionalité sont assez complémentaires. En effet, il faut faire attention aux individus atypiques pouvant nous pousser à rejeter l'hypothèse. On remarque assez bien ces individus grâce aux graphiques d'où l'intérêt de faire chacune de ces méthodes. Par exemple, on peut avoir un test significatif et une courbe de lissage ne s'apparentant pas à une droite horizontale uniquement à cause de certains individus.

Après avoir vérifié l'hypothèse des hasards proportionnels, intéressons nous à la vérification de l'hypothèse de log-linéarité.

2.3 Vérification de l'hypothèse de log-linéarité

Nous allons voir, dans cette partie, deux façons de vérifier l'hypothèse de log-linéarité définie en (2.2), permettant de déterminer la forme fonctionnelle de l'influence d'une covariable la plus adéquate pour chacune d'entre elles.

2.3.1 Résidus de martingales

Les résidus de martingales permettent de déterminer si une transformation de la covariable Z est nécessaire ou non.

Définition 2.3.1 (Résidus de martingales)**.**
Pour chaque observation, les résidus correspondent à la différence entre les valeurs observées d'une covariable et les valeurs prédites de celle-ci pendant un temps d'observation (Danieli, 2014, (5)). Ainsi, nous avons pour un individu i :

$$M_i(t) = N_i(t) - \Lambda_i(t),$$

avec :

- $N_i(t) \in \{0, 1\}$ qui correspond à un processus de dénombrement *i.e.* le nombre d'observations de l'évènement d'intérêt pour l'individu i à l'instant t. Par exemple, dans nos données `bfeed`, pour la femme 3, il n'y a pas d'observation de l'arrêt de l'allaitement donc $N_3(t) = 0$, en revanche on a observé l'arrêt de l'allaitement pour la femme 6 donc $N_6(t) = 1$;
- $\Lambda_i(t)$, appelée intensité cumulée (Amini, 2015, (2)), est donnée par :

$$\Lambda_i(t) = \int_0^t Y_i(u) e^{(Z^{(i)})^T \beta} h_0(u)\, \mathrm{d}u,$$

  où $Y_i(u)$ est une indicatrice permettant de savoir si l'individu i est encore à risque à la durée u. C'est-à-dire $Y_i(u) = 1$ si pour l'individu i nous n'observons pas l'évènement d'intérêt et qu'il n'y a pas de censure en t et $Y_i(u) = 0$ sinon.

Pour des données censurées à droite, nous nous intéressons à l'estimation pour $t = \infty$ suivante :

$$\begin{aligned}\hat{M}_i &= \hat{M}_i(\infty) \\ &= N_i(\infty) - \hat{\Lambda}_i(\infty) \\ &= N_i(\infty) - \int_0^\infty Y_i(u) e^{(Z^{(i)})^T \hat{\beta}} h_0(u)\, \mathrm{d}u,\end{aligned}$$

où $\hat{\beta}$ est l'estimateur de β et $N_i(\infty)$ correspond, ici, à la censure ou non de l'individu pendant toute la durée de l'étude *i.e.* $N_i(\infty)$ vaut 0 si l'individu i est censuré, 1 sinon.

Interprétation des résidus :

- Si $\hat{M}_i > 0$, l'individu i connait l'évènement d'intérêt plus tôt que ce qui était prévu ;
- Si $\hat{M}_i = 0$, l'individu i connait l'évènement d'intérêt en même temps que ce qui était prévu ;
- Si $\hat{M}_i < 0$, l'individu i est censuré ou connait l'évènement d'intérêt plus tard que ce qui était prévu.

Propriété 4.
Si la covariable Z a un effet de la forme $f(z)$ alors ses résidus sont approximativement proportionnels à $f(z)$. Ainsi, on a $\mathbb{E}(M|Z) \approx cf(z)$, avec c dépendant du nombre d'individus censurés (Quantin, 2018, (1)).

Représentation des résidus de martingale

Pour trouver la forme fonctionnelle adéquate, on peut représenter les résidus de martingales du modèle sans covariable en fonction des valeurs de chaque covariable (Quantin, 2018, (1)). Une tendance lissée des résidus, correspondant à $\mathbb{E}(M|Z)$, est ajoutée à cette représentation afin de mieux identifier la forme fonctionnelle adéquate. Si la tendance lissée des résidus est une droite alors la forme fonctionnelle adéquate est linéaire puisque $\mathbb{E}(M|Z) \approx cf(z)$ où $f(z)$ correspond à une droite.

Application avec les données `bfeed` : Pour la covariable `yschool` nous obtenons la représentation suivante :

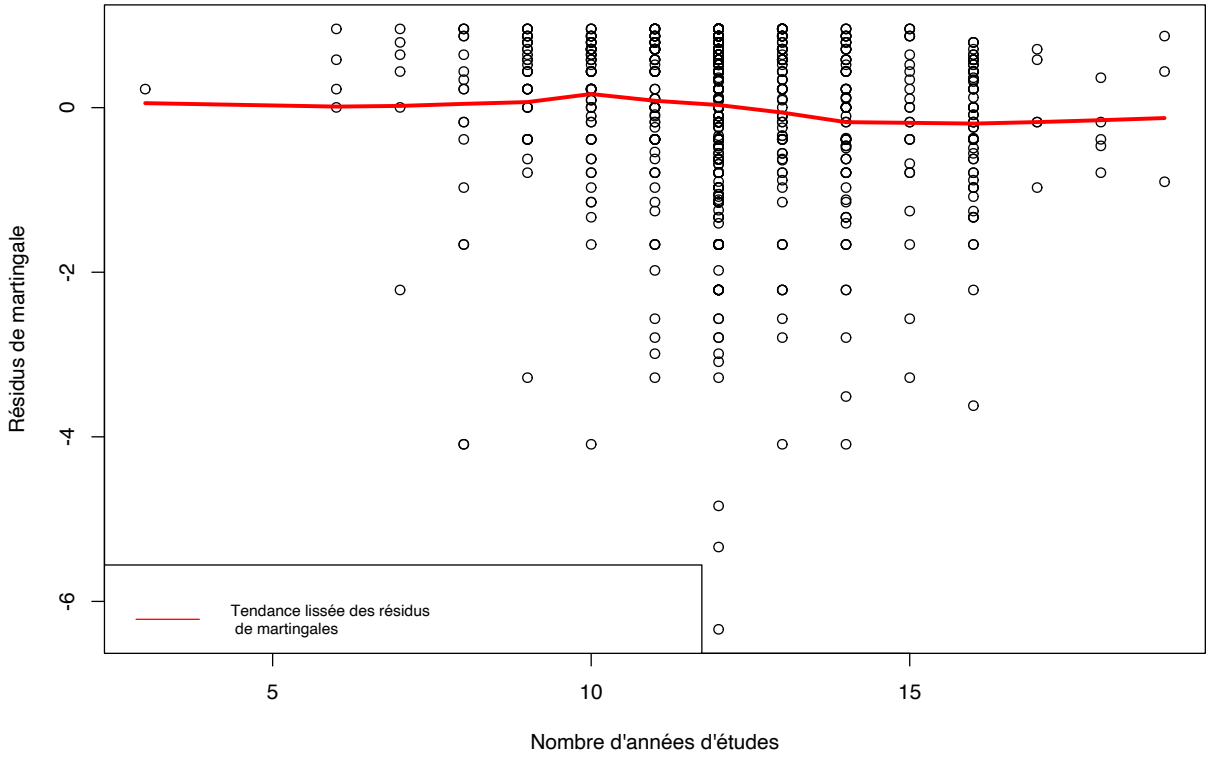

FIGURE 2.4 – Représentation des résidus de martingales de la covariable `yschool` en fonction du nombre d'années d'études.

La tendance lissée des résidus de martingales (en rouge), est ici similaire à une droite. Ainsi, la forme fonctionnelle adéquate pour les résidus de martingales de la covariable `yschool` est linéaire. Il y a donc une relation linéaire, mais aussi log-linéaire entre le nombre d'années d'études et la durée d'allaitement des femmes. De ce fait, on peut dire que la covariable `yschool` vérifie l'hypothèse de log-linéarité.

Néanmoins cette méthode ne permet pas de prendre en compte de possibles corrélations entre les covariables et ceci peut conduire à des formes fonctionnelles fausses. C'est pourquoi nous allons voir une deuxième méthode permettant d'avoir des formes fonctionnelles flexibles étant plus intéressantes en pratique.

2.3.2 Splines de lissage

L'utilisation des splines de lissage va permettre d'avoir une plus grande flexibilité sur les formes fonctionnelles en modélisant un effet non-linéaire d'une covariable sur le logarithme de la fonction de hasard. Dans cette partie nous allons voir comment vérifier l'hypothèse de log-linéarité du modèle de Cox à l'aide des splines.

Les splines sont des fonctions définies par morceaux par des polynômes. Elles peuvent être utilisées pour lisser des données expérimentales ou statistiques. On parle alors de "Smoothing splines". Cela va alors permettre d'utiliser dans le modèle une forme fonctionnelle plus flexible qu'une relation linéaire en considérant un polynôme en la covariable choisie.

Spline de lissage

Afin de déterminer ce qu'est une spline de lissage, il nous faut tout d'abord définir les B-splines. Un B-spline ou Basis-spline est une fonction polynomiale par morceaux et une combinaison de B-splines va permettre de donner une courbe lissée pour forme fonctionnelle dans le modèle (Amini, 2015, (2)).

Définition 2.3.2 (B-spline)**.**
Soit une suite de points $x_0 \leq ... \leq x_n$, appelés noeuds.
Selon Pansu (2014, (6)), une fonction B-spline de degrès k entre deux noeuds est la fonction $B_{i,k}$ définie par la relation de récurrence, avec $x \in [x_i, x_{i+k}]$:

$$B_{i,k}(x) = \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \Big(1 - \frac{x - x_{i+1}}{x_{i+k+1} - x_{i+1}}\Big) B_{i+1,k-1}(x),$$

et on a :

$$B_{i,0}(x) = B_i(x) = \begin{cases} 1 \text{ si } x \in [x_i, x_{i+1}]. \\ 0 \text{ sinon.} \end{cases}$$

Maintenant que nous avons défini les B-splines, nous pouvons voir comment se caractérise la courbe lissée déterminée par ces derniers.

Définition 2.3.3 (Courbe lissée)**.**
Soient les points $p_0, ..., p_m$, appelés points de contrôle, formant un polygone de contrôle.
Selon Amini (2015, (2)), une courbe de lissage par splines de degré n est la fonction S, composée de fonctions B-splines de degré k avec $k + m + 1 = n$, telle que :

$$S(x) = \sum_{j=1}^{m} B_{j,k}(x) p_j.$$

Application sur le modèle de Cox

Cette méthode est utilisée dans le modèle de regression de Cox pour vérifier l'hypothèse de log-linéarité. Pour une covariable Z_i on peut utiliser un spline de pénalité $S(Z_i)$ en supposant que la log-linéarité est vraie pour les covariables restantes. On va alors modifier légèrement le modèle de la façon suivante pour tout $t \geq 0$:

$$h(t|Z) = h_0(t) \exp(\beta_1 Z_1 + ... + \beta_{i-1} Z_{i-1} + S(Z_i) + \beta_{i+1} Z_{i+1} + ... + \beta_p Z_p)$$

On regarde ensuite si l'estimation avec spline, *i.e.* la courbe de splines lissée, est linéaire dans l'ensemble. Si cela est le cas alors l'hypothèse de log-linéarité est vérifiée pour la covariable en jeu. Sinon l'effet de cette covariable n'est pas log-linéaire et on peut transformer celle-ci à l'aide de la forme fonctionnelle trouvée afin de vérifier l'hypothèse de log-linearité.

Application sur les données `bfeed` : Dans l'exemple ci-dessous, nous allons représenter graphiquement, à l'aide des splines de lissage, l'effet de la covariable `yschool` sur le risque relatif h, en choisissant 12 ans comme valeur de référence.

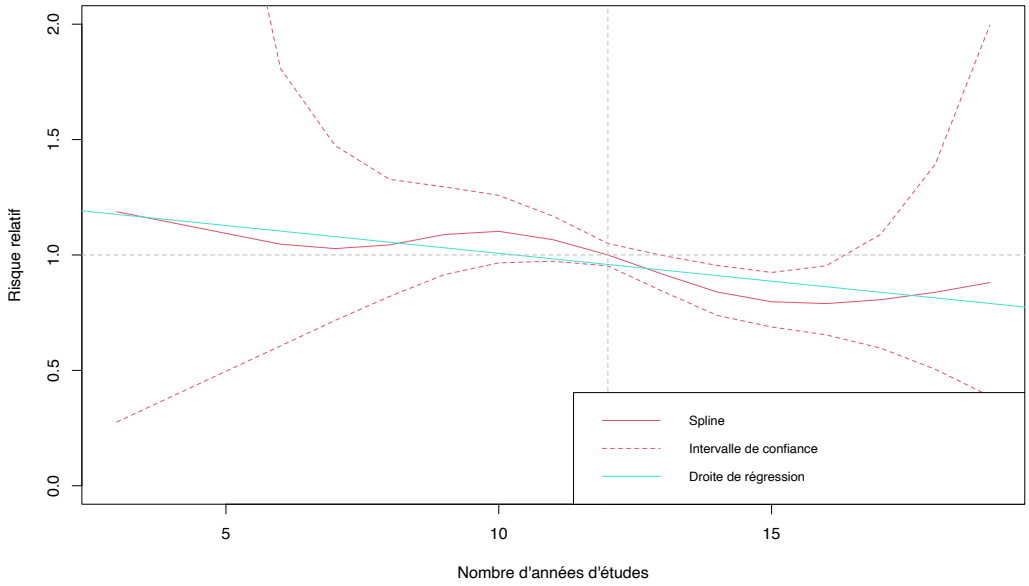

FIGURE 2.5 – Représentation du risque relatif lié au nombre d'années d'études (référence : 12 ans).

Nous pouvons observer sur ce graphique que la courbe de spline du risque relatif est linéaire et décroissante dans l'ensemble. On peut donc dire que le logarithme du risque relatif est aussi globalement linéaire et par conséquent, `yschool` vérifie l'hypothèse de log-linéarité. Ce graphique nous montre, de plus, que pour le niveau d'études de référence, qui est de 12 ans ici, nous avons bien un risque relatif valant 1.

Ainsi, nous avons le risque relatif entre une femme ayant fait 15 ans d'études et une autre ayant fait 12 ans d'études qui est d'environ -0.2. C'est-à-dire qu'une femme ayant fait 3 ans d'études supplémentaires a un risque d'arrêter d'allaiter qui diminue d'environ 0.2. Ici, nous avons choisi de prendre 12 années comme référence car ce niveau d'études correspond au niveau médian. Cependant, nous aurions pu prendre un autre niveau d'études comme référence, cela aurait simplement décalé la courbe spline de sorte qu'elle passe par le point d'abscisse le niveau d'études de référence et d'ordonnée 1, elle aura donc toujours la même allure quelque soit le niveau d'études de référence. Ainsi, l'interprétation des risques relatifs diffèrent peu entre les différents points de référence mais reste globalement la même.

Nous venons de voir comment vérifier les hypothèses du modèle de Cox et que l'on peut adapter la forme fonctionnelle afin de vérifier l'hypothèse de log-linéarité. Sur les données `bfeed`, nous avons vu que les covariables `yschool` et `smoke` ne vérifient pas l'hypothèse des hasards proportionnels. De plus, nous avons étudié les covariables `ybirth` et `race` du modèle (voir le code en annexe). La covariable `ybirth` vérifie l'hypothèse de proportionnalité alors que la covariable `race` ne la vérifie pas. L'hypothèse des hasards proportionnels étant plus difficile à vérifier puisque les variables peuvent être dépendantes du temps, nous verrons dans le prochain chapitre comment vérifier cette hypothèse pour des covariables temps-dépendantes.

CHAPITRE 3

EXTENSIONS DU MODÈLE DE COX

3.1 Covariables dépendantes du temps

Dans ce chapitre, nous allons voir comment appliquer le modèle de Cox sur des covariables dépendantes du temps, bien que l'hypothèse des hasards proportionnels ne soit pas vérifiée. Pour cela, nous verrons deux méthodes utilisées pour remédier à ce problème.

Définition 3.1.1 (Covariable temps-dépendante)**.** Une covariable temps-dépendante est une covariable dont l'effet ou dont la valeur varie au cours du temps.

Rejet de l'hypothèse des hasards proportionnels

Avec des covariables dont l'effet sur les rapports des hasards dépend du temps on a :

$$\frac{h(t|Z^{(i)})}{h(t|Z^{(j)})} = \exp((Z^{(i)} - Z^{(j)})^T \beta(t)).$$

On rejette alors l'hypothèse de proportionnalité.
Pour utiliser le modèle de Cox sur des covariables temps-dépendantes, il faut modifier le modèle. Pour ce faire, nous étudierons deux méthodes possibles :

- La stratification des covariables ;
- Le partitionnement du temps.

3.2 Stratification du modèle

Lorsque l'hypothèse des risques proportionnels n'est pas vérifiée, on peut stratifier le modèle afin d'étendre le modèle de Cox et de pouvoir l'appliquer sur des sous-échantillons (strates) de variables.

La stratification du modèle de Cox sur les covariables peut être effectuée facilement car celui-ci n'estime pas la fonction de risque de base. Pour ceci, on définit une fonction de risque de base pour chaque sous-échantillon (strate) en gardant le même risque relatif de chaque variable dans chacune des strates. Ainsi, on a pour une strate k :

$$h_k(t|Z) = h_{0,k}(t)\exp(Z^T\beta).$$

L'hypothèse des risques proportionnels pouvant être tenable sur des strates de taille n_k, on peut estimer les vraisemblances partielles de ceux-ci comme suit :

$$L_k^*(\beta) = \prod_{i=1}^{n_k} \frac{h_{0,k}(t_i)\exp(\beta_i Z_i)}{\sum_{l\in R_i} h_{0,k}(t_i)\exp(\beta_l Z_l)} = \prod_{i=1}^{n_k} \frac{\exp(\beta_i Z_i)}{\sum_{l\in G_i}\exp(\beta_l Z_l)}, \tag{3.1}$$

avec R_i le groupe de sujets à risque juste avant t_i.
Ce qui correspond à la vraisemblance partielle de Breslow définie en (1.3).
Finalement, la vraisemblance partielle du modèle de Cox stratifié avec K strates est (Giorgi, 2013, (4)) :

$$L^*(\beta) = \prod_{k=1}^{K} L_k^*(\beta).$$

La stratification de modèle permet alors de faire un ajustement naturel pour une covariable et ne repose plus sur l'hypothèse de proportionnalité puisque l'on récupère un modèle de Cox par strate (Giorgi, 2013, (4)). Néanmoins, on ne peut pas faire une estimation de l'importance de l'effet de chaque strate, de plus, il y a une perte de précision dans l'estimation des coefficients du modèle et plus il y a de strates, moins l'analyse est puissante.

Application avec les données `bfeed` : Nous avons vu dans la partie de vérification des hypothèses que la covariable `yschool` ne vérifiait pas l'hypothèse des hasards proportionnels. Partitionnons alors cette covariable afin de pouvoir étendre notre modèle de Cox.

Pour cela, nous déterminons les strates de façon empirique et nous obtenons deux intervalles : $\left[3, 14\right]$ et $\left[15, 19\right]$. Nous réalisons alors un test du log-rank qui nous permet de tester si les courbes de survie des deux strates que l'on vient de créer sont les mêmes ou non. Pour `yschool` nous avons une p-value pour le test du log-rank égale à 0.06. Ici, on ne rejettera pas l'hypothèse $\mathcal{H}_0$ selon laquelle les deux courbes de survie sont les mêmes, au risque de 5% mais on remarque qu'au risque de 7% on pourrait la rejeter.

On peut alors penser que les survies des deux groupes sont tout de même différentes. On pourra ensuite étendre le modèle de Cox avec cette covariable stratifiée.

Regardons également une représentation de ces courbes de survie.

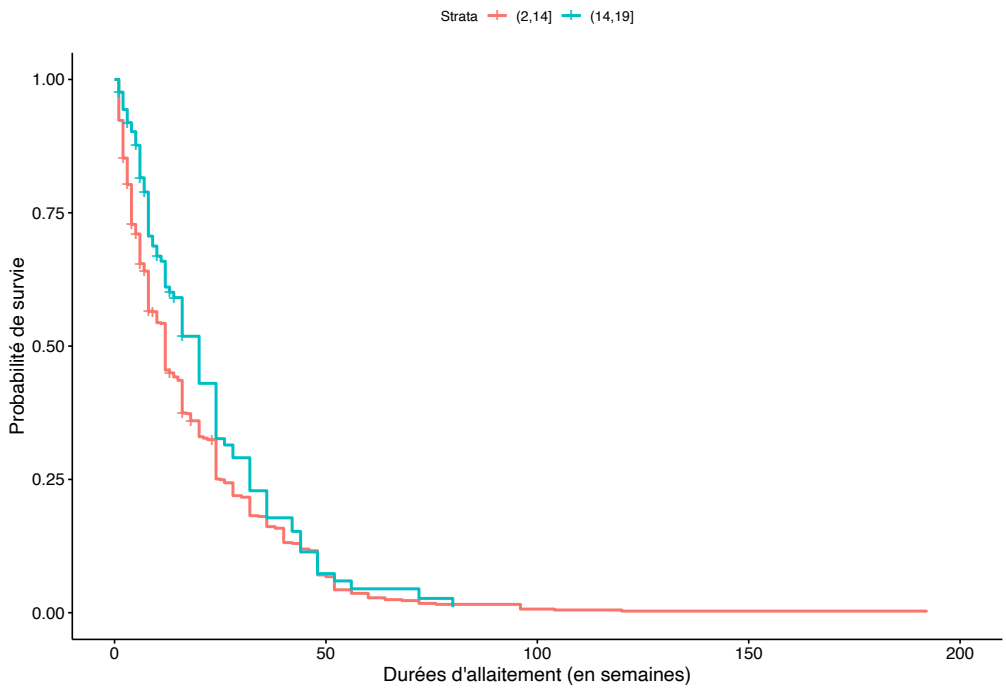

FIGURE 3.1 – Représentation des courbes de survie des femmes des deux strates de la covariable `yschool`.

On remarque que la courbe de survie des femmes ayant un cursus scolaire de plus de 14 ans est au dessus de la courbe de survie des femmes ayant réalisé au maximum 14 ans d'études. Ceci signifie que les femmes ayant fait moins de 14 ans d'études ont un risque d'arrêter d'allaiter plus élevé que les autres. De plus, dans ce modèle stratifié, on remarque que les femmes ayant fait plus de 14 ans d'études ont un risque d'arrêt d'allaitement 21% plus faible que celui des femmes ayant fait moins de 14 années d'études. Ici, on ne peut pas différencier le risque relatif d'une femme ayant fait 3 ans d'études par rapport à une autre qui en a fait 14, puisqu'elles sont les deux dans la première strate. De même, on ne peut pas différencier une femme ayant fait 15 ans d'études à une autre en ayant 19.

Si on regarde le risque relatif d'une femme ayant fait 5 ans d'études par rapport à une femme en ayant fait 15, on remarque qu'il est un peu différent du risque relatif entre les deux strates du modèle de Cox stratifié, puisqu'il est de $0.95107^{10} = 0.6055149$, c'est à dire qu'une femme ayant fait 15 ans d'études a 39% moins de risque d'arrêter d'allaiter, tandis que le risque entre les deux strates est de 0.79229, soit 21% de risque en moins d'arrêter d'allaiter. Cependant, le risque d'arrêter d'allaiter reste toujours plus faible chez les femmes ayant fait de longues études que chez celles ayant arrêter les études plus tôt.

Cette stratification nous permet de comparer uniquement les coefficients estimés des deux strates de la covariable `yschool` tandis qu'avec le modèle de Cox initial nous pouvions mesurer le risque exact entre les femmes de différents niveaux d'études. Par conséquent, l'interprétation des résultats avec la stratification reste moins précise que si l'on avait pu conserver la covariable quantitative.

En résumé, la stratification permet de contourner la non-vérification des hypothèses du modèle de Cox de sorte à ne pas sur-interpréter des résultats concernant des covariables ne vérifiant pas les hypothèses. Néanmoins, elle engendre une perte d'information.

3.3 Partitionnement du temps

Quand l'hypothèse des risques proportionnels n'est pas vérifiée, on peut aussi partitionner le temps en plusieurs intervalles et on regarde ensuite les résidus de Schoenfeld pour voir si sur chaque intervalle les hasards proportionnels sont vérifiés. Pour ce faire, il faut pour chaque individu générer une observation par intervalle.

Il est possible de cette manière d'approcher la forme fonctionnelle de $\beta(t)$ en utilisant des coefficients différents par intervalles de temps pour pouvoir mieux appréhender leur temps-dépendance. Cela revient, ainsi, à faire une stratification sur le temps.

Afin de vérifier l'hypothèse des hasards proportionnels à l'aide des résidus de Schoenfeld, nous pouvons supposer que le coefficient $\beta(t)$ est constant par morceaux.

On suppose $0 < t_{k+1} < t_k$ pour tout $k \in \{1, ..., K\}$.

Si $\beta(t)$ est constant par morceaux on a (Xu, 2018, (7)) :

$$\beta(t) = \sum_{k=1}^{K} \beta_k I_{[t_{k-1}, t_k[}(t),$$

où $I_{[t_{k-1}, t_k[}(t)$ vaut 1 si t est dans l'intervalle $[t_{k-1}, t_k[$ et 0 sinon.

Application avec les données `bfeed` : Nous avons vu dans le paragraphe 2.2.2 que la covariable `smoke` ne vérifiait pas l'hypothèse des hasards proportionnels, elle est donc dépendante du temps. Afin de réaliser un modèle de Cox contenant cette covariable, nous souhaitons réaliser une stratification du temps et pour cela nous supposons que $\beta(t)$ est constant par morceaux.
Afin de réaliser une bonne stratification, il nous faudra déterminer de façon empirique des intervalles de temps adéquats. On décide de se restreindre à 3 intervalles de temps car si on choisit trop d'intervalles, on risque d'avoir des erreurs trop importantes dues au trop grand nombre de coefficients.

Choisissons maintenant les bornes des intervalles.

Afin d'obtenir le meilleur partitionnement possible on décide de regarder les quantiles de la covariable et on remarque que 95% des observations se situent avant 48 semaines. On ne fera donc pas d'intervalles de temps au delà de 48 semaines puisqu'il n'y aurait que trop peu de valeurs dans le dernier intervalle de la stratification. Nous fixons donc une première borne à la médiane, qui est égale à 10.

Sachant que nous avons $[0; 10]$ comme premier intervalle, nous réalisons maintenant une boucle qui calcule la log-vraisemblance pour chaque modèle de Cox stratifié ayant un deuxième intervalle de temps compris entre 10 et 48. C'est-à-dire que pour chaque entier i on calcule la log-vraisemblance du modèle de Cox stratifié avec comme intervalles $[0; 10]$, $[10, i]$ et $[i, 192]$. Ainsi, on choisira la borne pour laquelle la log-vraisemblance est maximale.

Pour cela nous avons utilisé un code `R` basé sur l'algorithme suivant :

Algorithme 1 : Recherche de bornes d'intervalles optimales pour le partionnement du temps.

Initialisation :
$borne\_1 \leftarrow$ médiane(durées_allaitement);
$b \leftarrow$ vecteur de 10 à 48 par pas de 2;
$L^*(\beta) \leftarrow$ vecteur nul;
$n \leftarrow longueur(b)$;

Traitement :
pour *i allant de 1 à n* **faire**
  Calcul de la log-vraisemblance stratifiée défini en (3.1) pour les 3 strates $[0, borne\_1]$, $[borne\_1, b_i]$ et $[b_i, 192]$;
  $L^*(\beta)[i] \leftarrow$ log-vraisemblance stratifiée du modèle de Cox.
fin

$j \leftarrow$ indice de la valeur maximale du vecteur $L^*(\beta)$;
$borne\_2 \leftarrow b_j$;

Cet algorithme nous renvoit la valeur de $i = 22$ qui maximise la log-vraisemblance du modèle stratifié. Nous retenons alors les intervalles $[0; 10]$, $[10; 22]$ et $[22; 192]$ pour notre stratification de `smoke` par intervalles de temps.

Regardons maintenant la sortie `R` où l'on a appliqué une régression de Cox sur le modèle contenant la covariable stratifiée.

	coef	exp(coef)	se(coef)	z	p
`factor(race)2`	0.16318	1.17725	0.10348	1.577	0.11482
`factor(race)3`	0.27183	1.31236	0.09707	2.800	0.00510
`ybirth`	0.07424	1.07707	0.01797	4.132	3.6e-05
`yschool`	-0.05105	0.95023	0.01920	-2.660	0.00782
`smoke:strata:tgroup=1`	0.27281	1.31365	0.10392	2.625	0.00866
`smoke:strata:tgroup=2`	0.47529	1.60848	0.15430	3.080	0.00207
`smoke:strata:tgroup=3`	-0.04680	0.95428	0.16245	-0.288	0.77329
Test du rapport de vraisemblance = 44.55, on 7 df, p-value = 1.675e-07					
n= 1625, nombre d'évènements d'intérêt= 892					

TABLE 3.1 – Sortie R de la stratification de `smoke` par intervalles de temps

On remarque que les p-values des groupes "strata" sont très inférieures à 5% mis à part la dernière (ce qui est normal car cet intervalle ne contient pas beaucoup de valeurs). Donc pour chaque strate dont la p-value est inférieure à 0.05, on rejette l'hypothèse $\mathcal{H}_0$ selon laquelle la strate a un coefficient égal à 0, au risque de 5%. Ceci signifie que les coefficients associés à ces strates sont significativement différents de 0 au seuil de 5%. On peut donc penser que la covariable `smoke`, à travers ses strates, joue un rôle significatif dans le modèle de Cox stratifié.

De plus, on peut voir que le risque d'arrêter d'allaiter entre 0 et 10 semaines est 31% plus élevé pour les femmes fumeuses que pour les non-fumeuses et entre 10 et 22 semaines, il est 60% plus élevé pour les fumeuses que les non-fumeuses.

Regardons à présent la représentation des résidus et la stratification par ces intervalles de temps.

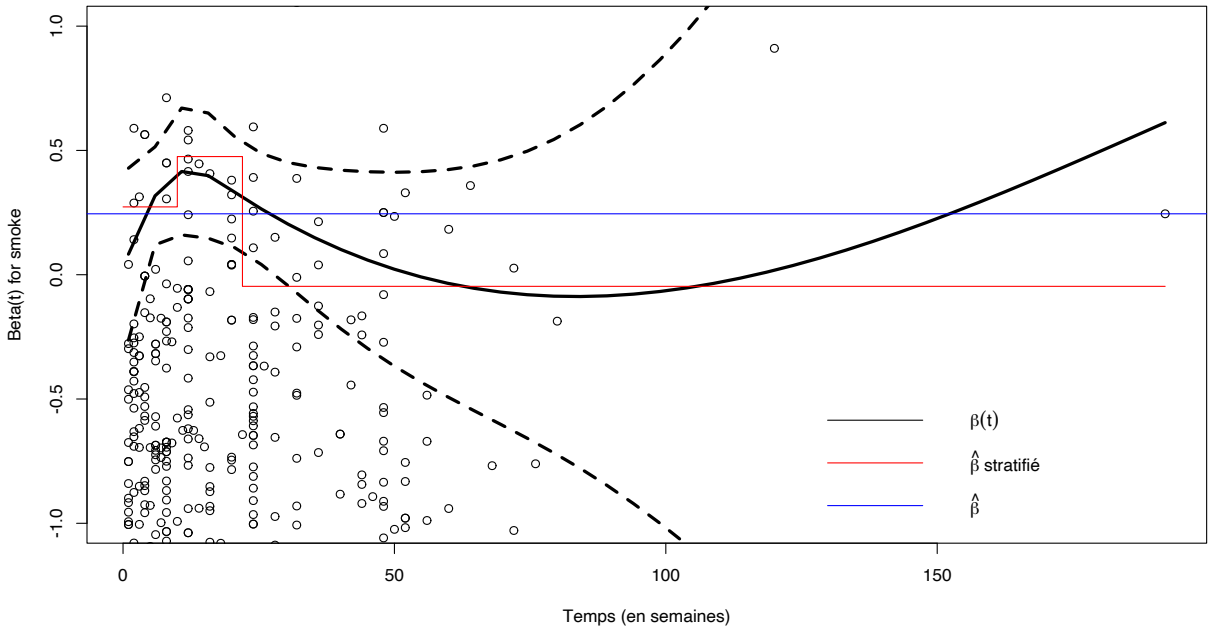

FIGURE 3.2 – Représentation des résidus de Schoenfeld et estimation des effets dépendants du temps de la covariable `smoke` par intervalles de temps.

On peut maintenant penser que la covariable `smoke` stratifiée par intervalle de temps est globalement indépendante du temps sauf sur le dernier intervalle.

Nous avons appliqué les extensions du modèle de Cox pour les covariables `yschool` et `smoke` ne vérifiant pas l'hypothèse de proportionnalité. Ainsi, nous avons vu que `smoke` et `yschool` vérifient globalement cette hypothèse par stratification. De ce fait, on peut retenir le modèle final avec les covariables `ybirth`, qui vérifie bien les hypothèses du modèle de Cox, `smoke` partitionnée par intervalles de temps, `yschool` stratifiée et `race`.

Voici la sortie R de l'estimation du modèle de Cox pour ce modèle final.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ybirth	0.06592	1.06814	0.01726	3.819	0.000134
strata.yschool(14,19]	-0.23480	0.79073	0.11296	-2.079	0.037653
factor(race)2	0.15599	1.16882	0.10364	1.505	0.132282
factor(race)3	0.31942	1.37632	0.09463	3.376	0.000737
smoke:strata(tgroup)tgroup=1	0.30201	1.35257	0.10257	2.944	0.003236
smoke:strata(tgroup)tgroup=2	0.50419	1.65564	0.15344	3.286	0.001016
smoke:strata(tgroup)tgroup=3	-0.01224	0.98784	0.16122	-0.076	0.939491
Test du rapport de vraisemblance = 42.02, on 7 df, p-value = 0.0000005					
n= 1625, nombre d'évènements d'intérêt= 892					

TABLE 3.2 – Sortie R de Coxph

Nous pouvons voir que ce modèle final est assez intéressant puisque les coefficients des covariables stratifiées sont assez semblables. En effet, le coefficient de la covariable `yschool` stratifiée dans ce modèle est quasiment égal au coefficient du modèle dans lequel la covariable `smoke` n'est pas partitionnée.

De même, les coefficients des strates de la covariable `smoke` sont également très semblables dans le modèle final et dans le modèle où la covariable `yschool` n'est pas stratifiée. Ceci signifie que l'on ne perd pas énormement d'information en stratifiant ces deux covariables, ce qui est assez satisfaisant.

De plus, les p-values sont inférieures à 5% mis à part pour `smoke:strate(tgroup)tgroup=3` comme nous l'avions vu précédemment et pour `factor(race)2`. Pour la covariable `race` il aurait fallu la stratifier car celle-ci ne vérifie pas l'hypothèse de proportionnalité. On ne peut alors pas bien analyser les résultats obtenus pour cette covariable puisque cela conduirait à surinterpréter des résultats concernant une covariable qui ne vérifie pas les hypothèses du modèle de Cox.

On peut donc penser que la covariable `smoke` et la covariable `yschool`, à travers leurs strates, jouent des rôles significatifs dans le modèle de Cox stratifié.

CONCLUSION

Ce travail, réalisé dans le cadre du projet de Master 1 de Biostatistique, avait pour but d'analyser l'utilisation du modèle de Cox avec la vérification des hypothèses de celui-ci au travers d'une application sur les données `bfeed`.

Pour ce faire, il a fallu définir dans un premier temps le modèle de Cox, les tests d'hypothèses associés et voir comment sélectionner le meilleur modèle possible. Dans un second temps, nous avons vu que l'hypothèse des hasards proportionnels peut être vérifiée graphiquement ou avec les résidus de Schoenfeld. Si cette hypothèse n'est pas vérifiée alors nous avons vu qu'il est possible d'adapter le modèle via la stratification du modèle ou le partitionnement du temps. De plus, nous avons vu que l'hypothèse de log-linéarité peut être vérifiée à l'aide des résidus de martingales ou des splines permettant de déterminer une forme fonctionnelle pour le modèle. Si la forme fonctionnelle n'est pas linéaire, il suffit d'appliquer cette dernière au modèle pour retrouver l'hypothèse de log-linéarité.

Enfin, lors de l'application sur les données `bfeed` nous avons vu que parmi les quatre covariables retenues dans le modèle, `yschool` et `smoke` ne vérifiaient pas l'hypothèse des hasards proportionnels. Il convenait alors de s'intéresser aux extensions du modèle de Cox pour ces covariables. L'idée était de stratifier ces covariables ou de partitionner le temps pour chacune afin de voir si elles vérifiaient cette hypothèse sur un modèle stratifié. En conclusion, les covariables `smoke` et `yschool` vérifient globalement cette hypothèse avec des modèles stratifiés. On a également vu que le modèle contenant les covariables stratifiées est intéressant car on ne perd pas beaucoup d'informations. On peut alors étudier ce modèle avec des covariables stratifiées. Dans ce rapport, nous avons choisi de ne pas traiter la covariable `race` en détail de sorte à ne pas alourdir notre étude et également par manque d'information sur celle-ci. Nous aurions pu aller plus loin avec cette covariable en la stratifiant de manière similaire à la covariable `smoke`.

BIBLIOGRAPHIE

[1] Simon QUANTIN. Rapport de l'INSEE, *Modèles semi-paramétriques de survie en temps continu sous R*. (2018).
[https://www.insee.fr/fr/statistiques/3695681]

[2] Zaki AMINI. Rapport de thèse, *Log-linearity for Cox's regression model*. (2015).
[https://www.duo.uio.no/bitstream/handle/10852/45377/thesis_zaki.pdf?sequence=15]

[3] André BERCHTOLD. Cours de Master, *Données longitudinales et modèles de survie*. (2014).
[https://andreberchtold.com/UNIGE/survie/4_Modele_Cox.pdf]

[4] Roch GIORGI. Cours Faculté de Médecine, *Extensions du modèle de Cox, Variables dépendantes du temps, Effets non linéaires*. (2013).
[https://sesstim.univ-amu.fr/sites/default/files/ressources_pedagogiques/extensions-cox-rg.pdf]

[5] Coraline DANIELI. Rapport de thèse, *Contributions méthodologiques à l'estimation de la survie nette*. (2014).
[https://tel.archives-ouvertes.fr/tel-01199181/file/TH2014DanieliCoraline.pdf]

[6] Pierre PANSU. Cours de Master, *Courbes B-splines*. (2004).
[https://www.imo.universite-paris-saclay.fr/~pansu/web_maitrise/bsplines.pdf]

[7] Ronghui XU. Cours de Biostatistique, *Assessing the Fit of the Cox Model*. (2018).
[http://www.math.ucsd.edu/~rxu/math284/slect7.pdf]

[8] Frédéric BERTRAND et Myriam MAUMY. Cours de Master, *Choix du modèle*. (2008).
[https://irma.math.unistra.fr/~fbertran/enseignement/Master1_2010_2/Master2Cours3.pdf]

[9] Gilbert SAPORTA. Cours du CNAM, *Choix de modèles*. (2012).
[http://cedric.cnam.fr/~saporta/selectionmodeles.pdf]

[10] Michaël CHOUKROUN. Bulletin français d'actuariat, Vol. 8, n°16, *Le modèle additif D'AALEN, une alternative au modèle de Cox dans le cadre de la construction d'une loi de maintien en incapacité de travail.* (2008).
[http://www.ressources-actuarielles.net/EXT/IA/sitebfa.nsf/0/BC591BD5F7B07218C12574CB002C3FA2/$FILE/CHOUKROUN.pdf?OpenElement]

[11] Thi Mong Ngoc NGUYEN. Rapport de thèse, *Estimation récursive pour les modèles semi-paramétriques*. (2014).
[https://tel.archives-ouvertes.fr/tel-00938607/document]

[12] Norman Edward BRESLOW. Article dans *Biometrics*, 30, 89-99, *Covariance analysis of censored survival data*. (1974).
[https://www.jstor.org/stable/2529620?origin=crossref&seq=1]

[13] David COX. Article dans *Journal of the Royal Statistical Society*, Series B, 34, 187-220., *Regression models and life-tables.* . (1972).
[http://www.stat.cmu.edu/~ryantibs/journalclub/cox_1972.pdf]

[14] Bradley EFRON. Article dans *Journal of the American Statistical Association*, 72 (359) : 557–565, *The efficiency of cox's likelihood function for censored data*. (1977).
[https://www.jstor.org/stable/i314242?refreqid=excelsior%3A167ea67f8fda03b28298940a38090a1f]

[15] Klein and Moeschberger. Livre, *Survival Analysis Techniques for Censored and truncated data.* (1997).

[16] Patrick BREHENY. Cours, *Quantifying predictive accuracy in Cox models.* (2019).
[https://myweb.uiowa.edu/pbreheny/7210/f15/notes/11-19.pdf]

ANNEXE A

CODE R

Lien git pour les codes utilisés dans la partie *Présentation des données* ici.

Lien git pour les codes utilisés dans la partie *Séléction de variable et test d'hypothèse* ici.

Lien git pour les codes utilisés dans la partie *Extension du modèle de Cox* ici.

Lien git pour les codes utilisés dans la partie *Hypothèses et vérifications* ici.