Le modèle de Cox

Victor Fardel Ewen Gallic

30 janvier 2013

Faculté des Sciences économiques, UFR de mathématiques

- Développé par David Cox (1972)[1]
- Objectifs et utilisation :
  - étudier la probabilité de défaillance d'un individu en fonction de ses caractéristiques, à différents instants t,
  - principaux domaines d'application :
    - industrie,
    - médecine,
    - sciences actuarielles ;
- Difficultés :
  - nombre de données limitées,
  - censure et troncature.

Présentation des données

- Rossi (1980)[5] ;
- 432 entrées ;
- 10 variables.

	arrest	week	fin	age	race	wexp	mar	paro	prio	educ
1	1	20	0	27	1	0	0	1	3	3
2	1	17	0	18	1	0	0	1	8	4
3	1	25	0	19	0	1	0	1	13	3
4	0	52	1	23	1	1	1	1	1	5
5	0	52	0	19	0	1	0	1	3	3
6	0	52	0	24	1	1	0	0	2	4
	...	...	...	...	...	...	...	...	...	...

Plan

Définition (Fonction de survie)

- *Soit X_i le temps entre le début de l'observation et la défaillance ;*
- *Les X_i sont i.i.d. de fonction de densité $f(x)$ et de fonction de survie*

$$x > 0, \quad S(x) = \mathbb{P}(X > x) \quad \textit{(v.a. discrète)}$$

$$x > 0, \quad S(x) = 1 - \int_0^x f(u)du \quad \textit{(v.a. continue)}$$

Dans notre exemple, X_i représente le temps entre la sortie de prison et la prochaine arrestation.

Définition (Fonction de survie résiduelle)

$$\forall x \geq 0, \quad S_t(x) = \mathbb{P}(X - t > x \mid X > t) = \frac{S(t+x)}{S(t)}$$

▸ Démo.

Dans l'exemple, la survie résiduelle correspond à la probabilité que l'individu ne soit pas de nouveau arrêté sachant qu'à l'instant t, il n'a pas encore récidivé.

Plan

Définition (Taux de défaillance)

$$\forall x > 0 \quad \alpha(x) = \lim_{h \downarrow 0} \frac{1}{h} \mathbb{P}\left(x \leq X < x + h \mid X \geq x\right)$$
$$= \frac{-S'(x)}{S(x)} = \frac{f(x)}{S(x)}$$

En effet, on a :

$$f(x) = \lim_{h \downarrow 0} \frac{1}{h} \mathbb{P}\left(x \leq X < x + h\right) = F'(x) = -S'(x)$$

Appliqué à notre jeu de données, il s'agit de la probabilité que l'individu se fasse arrêter entre les instants x et $x+h$ étant donné qu'à l'instant x il n'ait pas récidivé.

Plan

Définition (Censure)

On parle de censure C_i d'une donnée i si :

- *avant la fin de l'étude, on n'observe plus l'individu sans qu'il y ait eu de défaillance ;*
- *à la fin de l'étude, la défaillance n'a toujours pas été observée.*

Sinon, si une défaillance est observée, on qualifie la donnée de complète.
La réponse à la fin de l'étude pour un individu i est notée :

$$X_i^{obs} = \min(X_i, C_i)$$

Censure

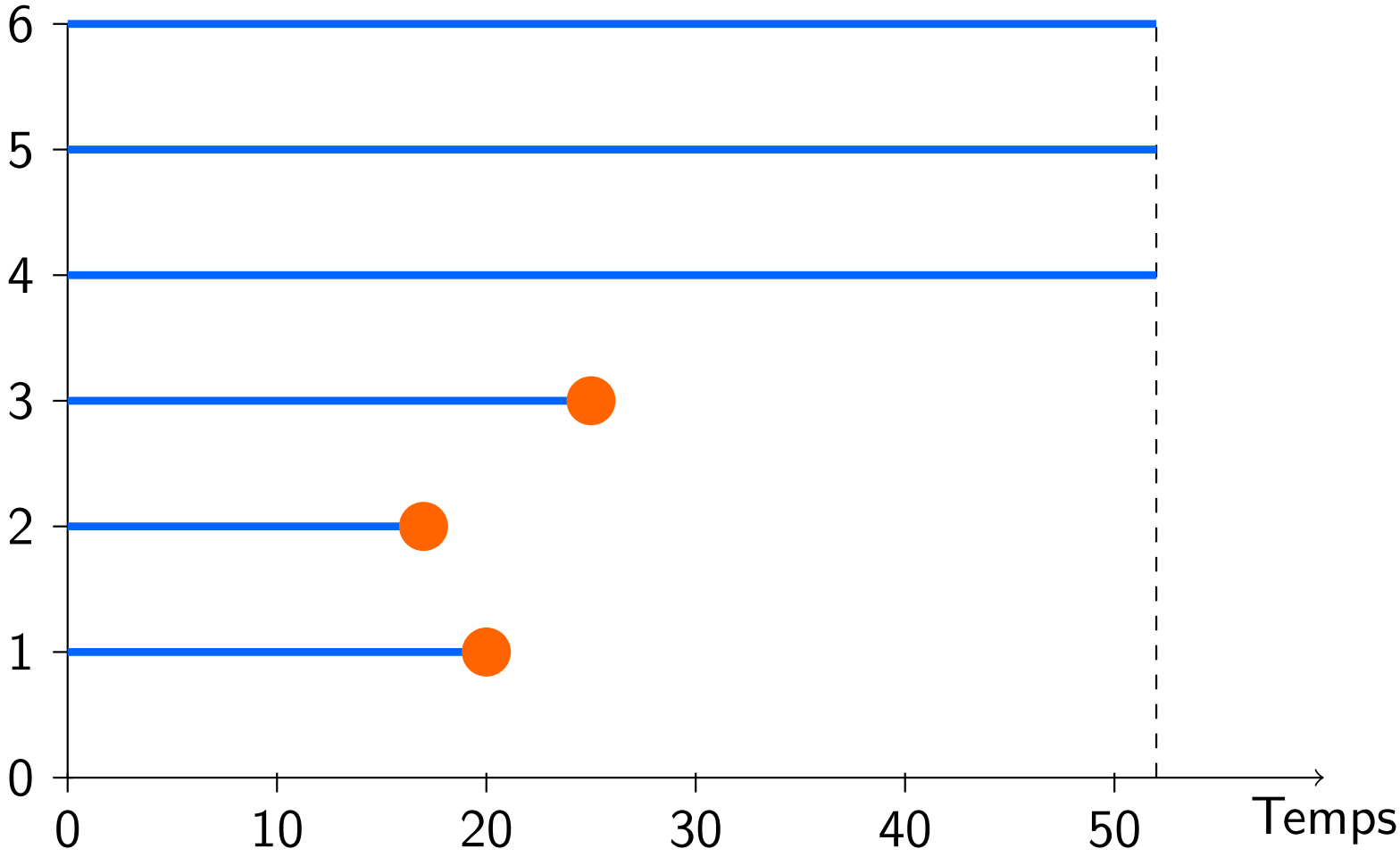

FIGURE 1: Illustration de la censure sur les données de prison.

Une indicatrice indique si la donnée i est censurée :

$$\delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$$

Dans le cas des prisonniers, si à l'issue de la 52^{e} semaine, l'individu n'a toujours pas été arrêté, la donnée est censurée (la variable *arrest* (δ) prend la valeur 0).

Plan

Définition (Processus de comptage des défaillances)

- T_i *la date de de défaillance pour l'individu* i ;
- $N(t)$ *la variable aléatoire indiquant le nombre de défaillances observées à l'instant* t. *Au début de l'étude, on a* $N(0) = 0$;
- $$Y_i(t) = \begin{cases} 1 \text{ si l'individu } i \text{ est encore à risque à } t \\ 0 \text{ sinon.} \end{cases}$$

Plan

Définition (écriture du modèle)

$$\alpha(t \mid \mathbf{Z}_i(t)) = \alpha_0(t) \exp(\mathbf{Z}_i^\top(t)\beta)$$

Avec :

- $\alpha_0(t)$ *le risque de base, une fonction non spécifiée (intensité en l'absence d'effet des covariables) ;*
- $\mathbf{Z}_i$ *le vecteur des covariables pour un individu* i *;*
- β *les mesures d'influence des covariables sur l'intensité.*

Plan

Hypothèse des risques proportionnels

Pour deux individus i et j, $\forall i \neq j$ on remarque :

$$\frac{\alpha(t \mid \mathbf{Z}_i)}{\alpha(t \mid \mathbf{Z}_j)} = \frac{\exp(\mathbf{Z}_i^\top \beta)}{\exp(\mathbf{Z}_j^\top \beta)} = \exp\left((\mathbf{Z}_i - \mathbf{Z}_j)^\top \beta\right)$$

- ne dépend que des individus i et j ;
- constant au cours du temps.

Plan

- On cherche à savoir la probabilité qu'un individu i avec des covariables $\mathbf{Z}_i$ ait un événement au moment T_j, sachant qu'un événement a lieu au moment T_j :

$$\mathbb{P}(\text{indiv. } i \text{ ait un évé. à } T_j \mid \text{un évé. à } T_j)$$

$$= \frac{\mathbb{P}(\text{indiv.} i \text{ ait un évé. à } T_j)}{\mathbb{P}(\text{un évé. à } T_j)};$$

- C'est la probabilité que l'individu i soit arrêté à la date T_j sachant qu'il y a eu une arrestation à la date T_j.

- Le numérateur est le taux de défaillance pour l'individu i au temps T_j ;
- le dénominateur est la somme des taux de défaillances pour l'ensemble des individus encore à risque à T_j ;
- La contribution de l'individu i à la vraisemblance est donc :

$$\begin{aligned}
\mathcal{L}_i(\beta) &= \frac{\alpha_i(T_i \mid Z_i)}{\sum_{j=1}^{n} Y_j(T_i)\alpha_j(T_j \mid Z_j)} \\
&= \frac{\alpha_0(T_i)\exp(\mathbf{Z}_i^\top(T_i)\beta)}{\sum_{j=1}^{n} Y_j(T_i)\alpha_0(T_i)\exp(\mathbf{Z}_i^\top(T_i)\beta)} \\
&= \frac{\exp(\mathbf{Z}_i^\top(T_i)\beta)}{\sum_{j=1}^{n} Y_j(T_i)\exp(\mathbf{Z}_i^\top(T_i)\beta)}
\end{aligned}$$

- La fonction de vraisemblance partielle s'écrit :

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^\top \beta)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\mathbf{Z}_j^\top \beta)} \right\}^{\delta_i},$$

  Avec T_i la date de défaillance de l'individu i.

- On en dérive l'expression du score partiel :

$$U_p(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \mathbf{Z}_j \exp(\mathbf{Z}_j^\top \beta)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\mathbf{Z}_j^\top \beta)} \right\}$$

▸ Démo.

Pour maximiser la vraisemblance partielle (soit maximiser le score partiel), et ainsi déterminer les paramètres β, il est fréquent d'utiliser l'algorithme de Newton-Raphson :

$$\beta_{k+1} = \beta_k + \mathcal{I}_n^{-1}(\beta_k)\, U_p(\beta_k),$$

avec

$$\mathcal{I}_n = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}_p(\beta)}{\partial\beta\partial\beta^\top}\right].$$

Il faut fournir une valeur β_0 initiale ainsi qu'un critère d'arrêt.

Les coefficients sont estimés à l'aide de l'algorithme de Newton-Raphson sous $\mathcal{R}$.

```
> mod <- coxph(Surv(week, arrest) ~ fin + age + race + wexp
    + mar + paro + prio, data=prison)
> print(mod)

         coef exp(coef) se(coef)      z      p
fin   -0.3794     0.684   0.1914 -1.983 0.0470
age   -0.0574     0.944   0.0220 -2.611 0.0090
race   0.3139     1.369   0.3080  1.019 0.3100
wexp  -0.1498     0.861   0.2122 -0.706 0.4800
mar   -0.4337     0.648   0.3819 -1.136 0.2600
paro  -0.0849     0.919   0.1958 -0.434 0.6600
prio   0.0915     1.096   0.0286  3.194 0.0014

Likelihood ratio test=33.3  on 7 df, p=2.36e-05  n= 432,
    number of events= 114
```

Plan

Théorème

- $\frac{1}{\sqrt{n}} U_p(\beta) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \Sigma)$;
- $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \Sigma^{-1})$;
- $\frac{1}{n}\mathcal{I}_n(\hat{\beta})$ *estimateur consistant de* Σ.

Test de significativité d'un coefficient :

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

Sous H_0, la statistique de test est de la forme :

$$\sqrt{n} \cdot \frac{\hat{\beta}}{\hat{\sigma}} \xrightarrow[n \to \infty]{\mathscr{L}} \mathcal{N}(0,1).$$

Un intervalle de confiance de niveau $1-\alpha$ est alors donné par :

$$\widehat{IC}_{1-\alpha}(\beta) = \left[\hat{\beta} \pm u_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

Application aux données

	coef	exp(coef)	se(coef)	z	p
fin	-0.38	0.68	0.19	-1.98	0.05
age	-0.06	0.94	0.02	-2.61	0.01
race	0.31	1.37	0.31	1.02	0.31
wexp	-0.15	0.86	0.21	-0.71	0.48
mar	-0.43	0.65	0.38	-1.14	0.26
paro	-0.08	0.92	0.20	-0.43	0.66
prio	0.09	1.10	0.03	3.19	0.00

Interprétation des coefficients

- On suppose que l'hypothèse des risques proportionnels tient ;

$$\frac{\alpha(t \mid Z_i)}{\alpha(t \mid Z_j)} = \exp\left((Z_i - Z_j)^\top \beta\right);$$

- Exemple de l'effet marginal d'un an supplémentaire :

$$\frac{\exp\left(\mathit{fin}_i \times \beta_{\mathit{fin}} + \mathit{age}_i \times \beta_{\mathit{age}} + \ldots + \mathit{prio}_i \times \beta_{\mathit{prio}}\right)}{\exp\left(\mathit{fin}_i \times \beta_{\mathit{fin}} + (\mathit{age}_i + 1) \times \beta_{\mathit{age}} + \ldots + \mathit{prio}_i \times \beta_{\mathit{prio}}\right)}$$

$$= \frac{\exp\left(\mathit{age}_i \times \beta_{\mathit{age}}\right)}{\exp\left((\mathit{age}_i + 1) \times \beta_{\mathit{age}}\right)} = \exp(\beta_{\mathit{age}}) = 0.94;$$

- Quel que soit t, chaque année de vie supplémentaire (*cet. par.*) multiplie le risque instantané par un facteur 0.94, soit une diminution de 6%.

Application aux données

	coef	exp(coef)	se(coef)	z	p
fin	-0.38	0.68	0.19	-1.98	0.05
age	-0.06	0.94	0.02	-2.61	0.01
race	0.31	1.37	0.31	1.02	0.31
wexp	-0.15	0.86	0.21	-0.71	0.48
mar	-0.43	0.65	0.38	-1.14	0.26
paro	-0.08	0.92	0.20	-0.43	0.66
prio	0.09	1.10	0.03	3.19	0.00

Interprétation des coefficients

- Recevoir une aide financière versus ne pas en recevoir, *cet. par*, diminue le risque instantané d'être arrêté de 32% ;
- La couleur de peau, le fait d'avoir eu un emploi à temps plein avant l'incarcération, le statut marital ou le fait d'avoir été libéré sous parole n'influent pas significativement sur le risque instantané ;
- Quel que soit t, chaque incarcération passée supplémentaire fait augmenter le risque de récidive de 10%, *cet. par.*

Test de nullité simultanée des coefficients

- Rapport de vraisemblance ;
- Test de Wald ;
- Score (ou logrank) ;
- Loi asymptotique : χ^2_p, avec p le nombre de paramètres du modèle.

```
Likelihood ratio test= 33.27  on 7 df,   p=2.362e-05
Wald test            = 32.11  on 7 df,   p=3.871e-05
Score (logrank) test = 33.53  on 7 df,   p=2.11e-05
```

Fonction de survie

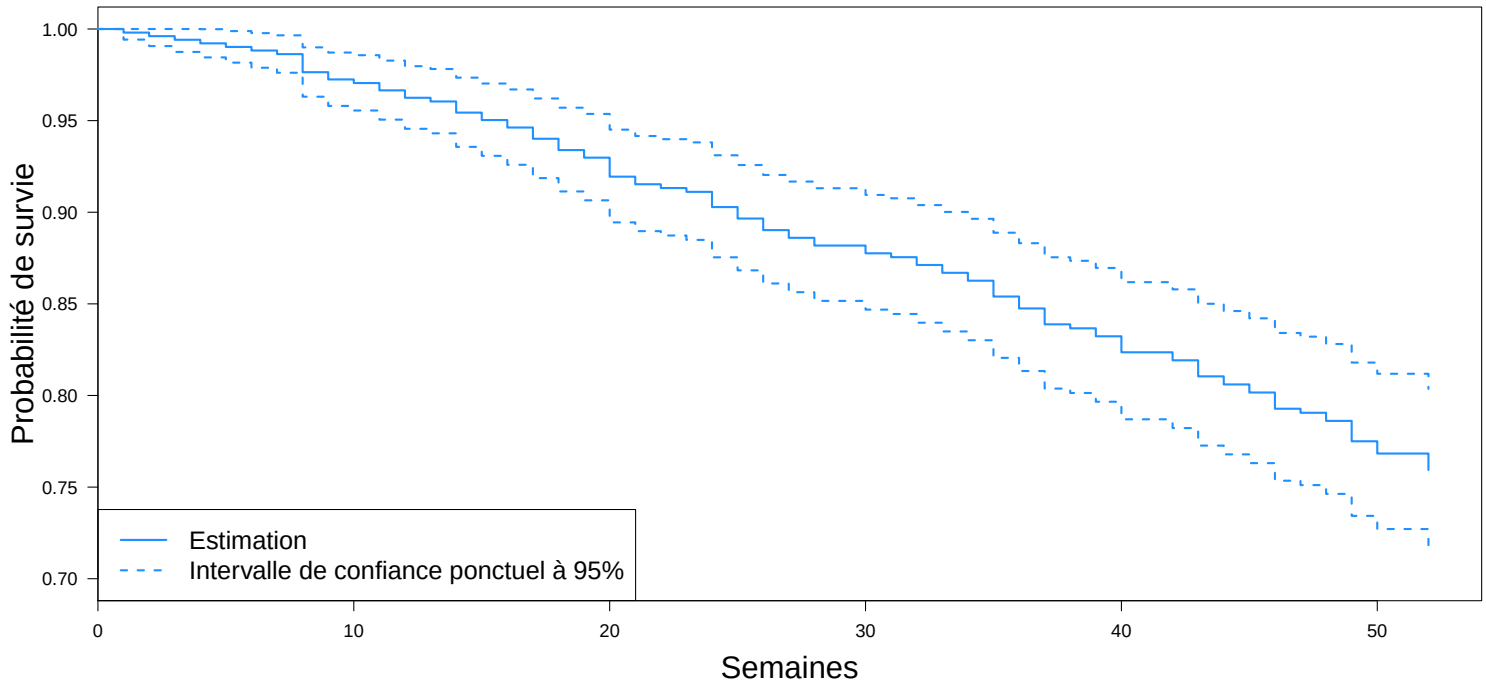

FIGURE 2: Estimation de la fonction de survie.

```
plot(survfit(mod), ylim=c(.7, 1))
```

Hypothèse des risques proportionnels

- Vérification de l'hypothèse des risques proportionnel à partir de tests :
  - numérique ;
  - graphique (ex : tracé des résidus de Shoenfeld normalisés).
- À tester sur un modèle avec des variables dont le coefficient est significativement non nul ;

Test numérique

Pour chaque variable, on cherche à s'assurer que le coefficient associé est stable au cours du temps.

$$\begin{cases} H_0 : \beta_j(t) = \beta_j \\ H_1 : \beta_j(t) \neq \beta_j. \end{cases}$$

▸ Détail.

Dans $\mathcal{R}$:

```
> mod2 <- coxph(Surv(week, arrest) ~ fin + age + prio, data
    = prison)
> cox.zph(mod2)
             rho    chisq       p
fin1    -0.00657 0.00507 0.9433
age     -0.20976 6.54147 0.0105
prio    -0.08004 0.77288 0.3793
GLOBAL        NA 7.13046 0.0679
```

Test graphique - Résidus de Schoenfeld

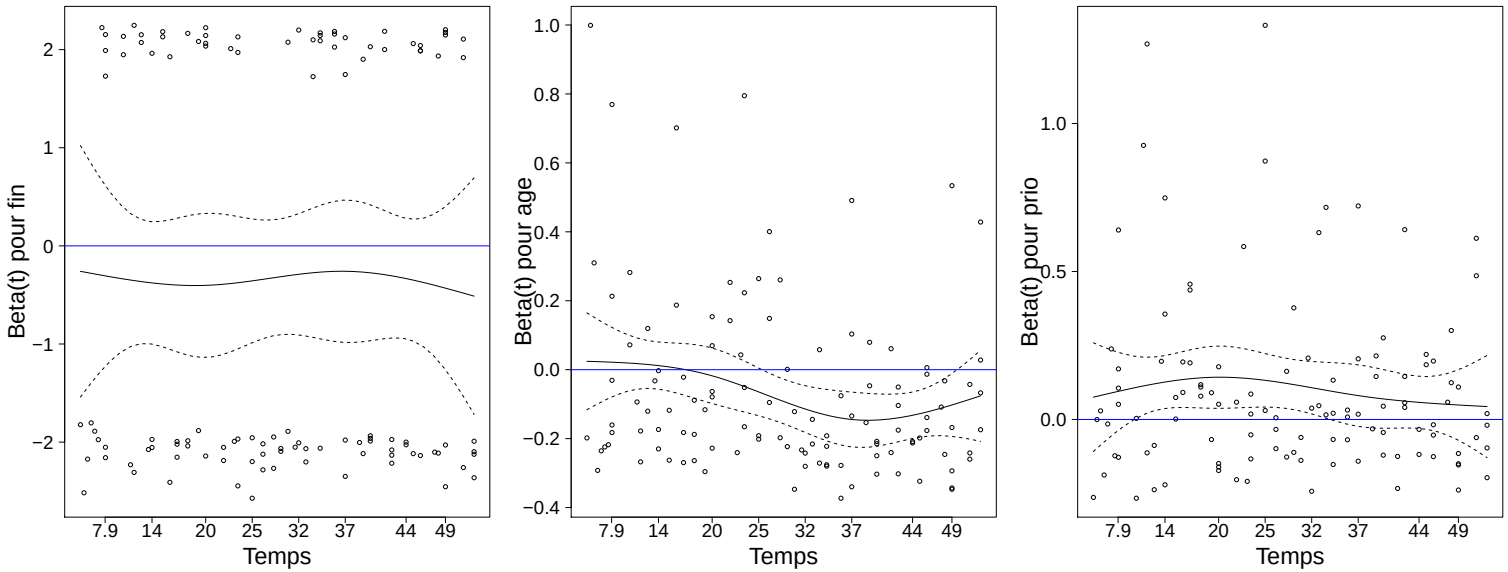

FIGURE 3: Résidus normalisés de Shoenfeld.

```
plot(cox.zph(mod2)[1]);abline(h=0,col="blue")
```

Test graphique - Résidus de martingale

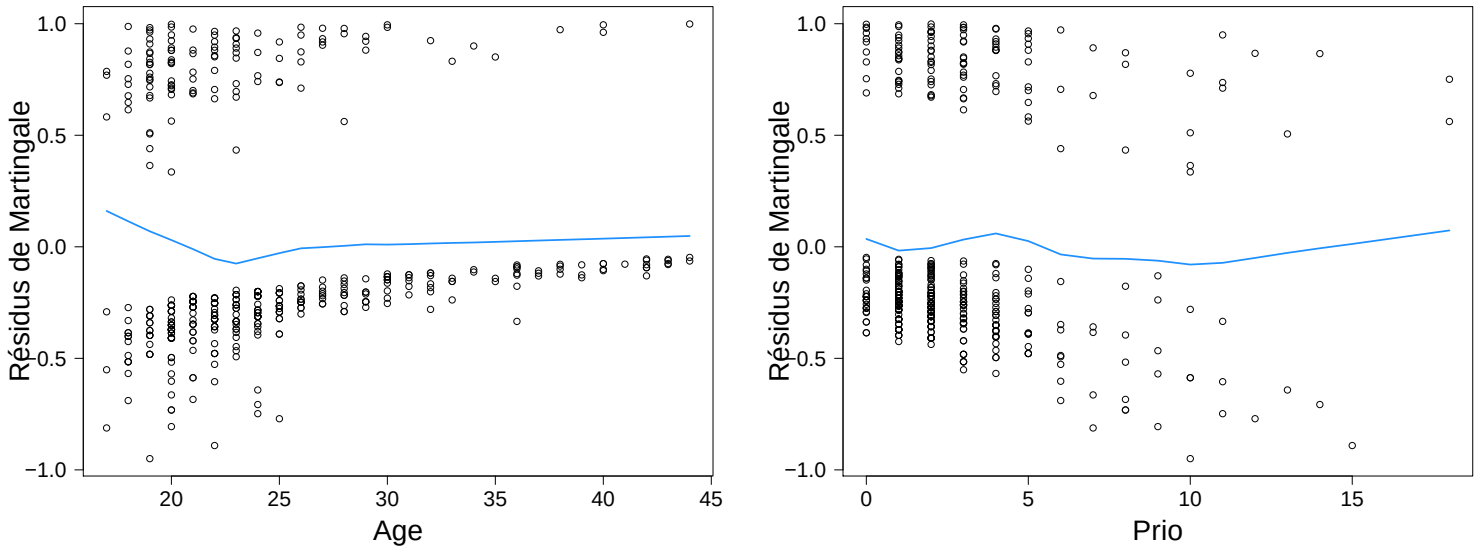

FIGURE 4: Résidus de martingale.

```
plot(residus.martingale~prison[,"age"],xlab="Age",ylab="
    Residus de Martingale")
lines(lowess(prison[,"age"], residus.martingale, iter=0),lwd
    =2,col="dodger blue")
```

Test graphique - Résidus de martingale

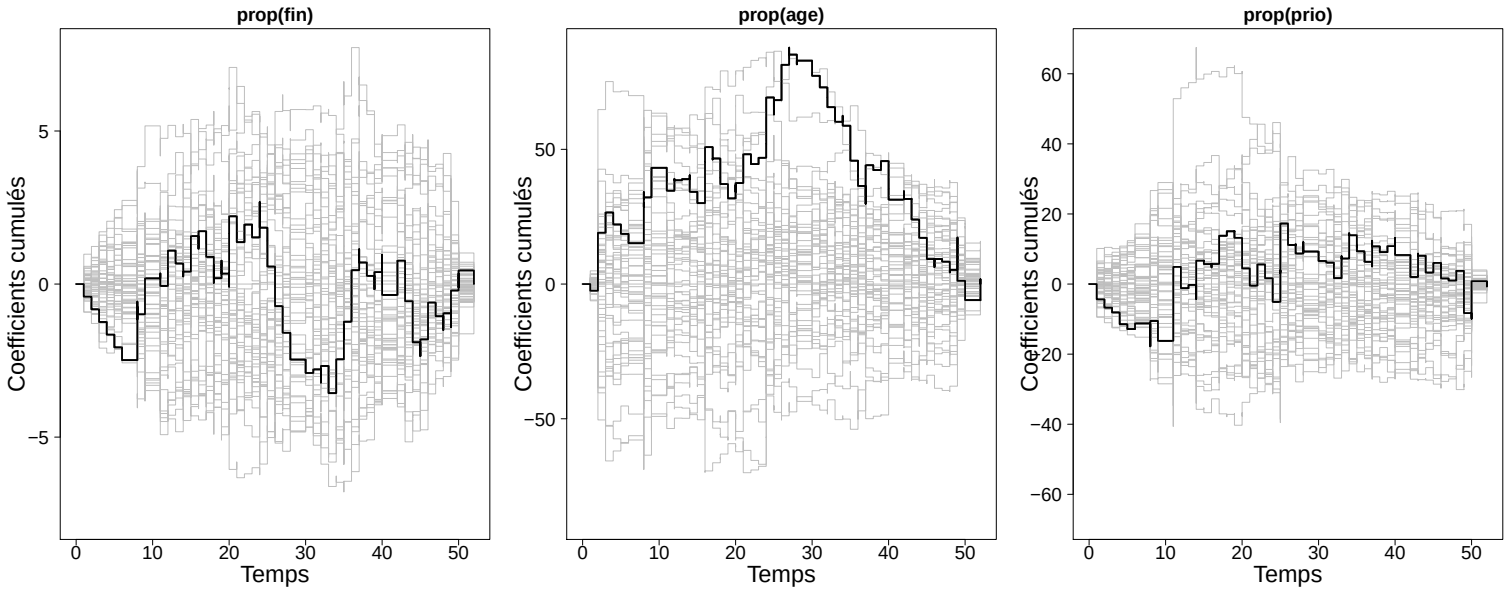

FIGURE 5: Résidus de martingale (Modèle Cox-Aalen).

```
op <- par(mfrow=c(1,3))
plot(cox.aalen(Surv(week, arrest) ~ prop(fin) + prop(age) +
    prop(prio), data = prison,n.sim=100),score=1)
par(op)
```

Pour aller plus loin...

- Interactions ;
- Stratification ;
- Covariables dépendantes du temps ;

Quelques références

D. R. Cox.

Regression models and life-tables.
Journal of the Royal Statistical Society. Series B (Methodological), 34(2) :pp. 187–220, 1972.

Jean-Fran¸cois Dupuy and James Ledoux.

Analyse de durées de vie. fiabilité.
Notes de cours, 2012.

John Fox.

Cox proportional-hazards regression for survival data the cox proportional-hazards model.
Most, 2008(June) :1–18, 2002.

Laurence Reboul.

Cours de durées de vie.
Notes de cours, 2011.

P.H. Rossi, R.A. Berk, and K.J. Lenihan.

Money, work, and crime : experimental evidence.
Quantitative studies in social relations. Academic Press, 1980.

Marie-Luce Taupin.

Durées de survie.
Notes de cours, 2011.

$$\begin{aligned}\mathbb{P}(X - t > x \mid X > t) &= \frac{\mathbb{P}(X - t > x \cap X > t)}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(X > x + t)}{\mathbb{P}(X > t)} \\ &= \frac{S(x + t)}{S(x)}\end{aligned}$$

▸ Retour fonction de survie résiduelle.

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^\top \beta)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\mathbf{Z}_j^\top \beta)} \right\}^{\delta_i}$$

$$\ell_p(\beta) := \log \mathcal{L}_p(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{Z}_i^\top \beta - \log \sum_{j=1}^{n} Y_j(T_i) \exp(\mathbf{Z}_i^\top \beta) \right\}$$

Le score partiel s'obtient en dérivant la log-vraisemblance partielle par rapport à β :

$$U_p(\beta) := \frac{\partial \ell_p(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \mathbf{Z}_j \exp(\mathbf{Z}_j^\top \beta)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\mathbf{Z}_j^\top \beta)} \right\}$$

▸ Retour vraisemblance partielle.

Statistique de test de l'hypothèse des risques proportionnels (1/3)

Le modèle de Cox s'écrit :

$$\alpha(t \mid \mathbf{Z}_i(t)) = \alpha_0(t) \exp(\mathbf{Z}_i^\top(t)\beta(t)).$$

On a considéré depuis le début que $\forall t, \beta(t) = \beta$ (risques proportionnels vérifiée). Pour chacune des $j = 1, \ldots, p$ covariables, on a donc :

$$\beta_j(t) = \beta_j + \theta_j g_j(t),$$

avec $g_j(t)$ une fonction connue.

Statistique de test de l'hypothèse des risques proportionnels (2/3)

$$\beta_j(t) = \beta_j + \theta_j g_j(t)$$

Pour vérifier l'hypothèse des risques proportionnels, on doit avoir

$$\forall j, \theta_j = 0.$$

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta_j \neq 0. \end{cases}$$

Statistique de test de l'hypothèse des risques proportionnels (3/3)

La statistique de test est la suivante (sous H_0) :

$$Z = U_2^\top(\widehat{\beta}, 0)\widehat{I_n^{22}}(\widehat{\beta}, 0)\, U_2(\widehat{\beta}, 0) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p,$$

- avec $U_2^\top(\widehat{\beta}, \theta)$ la dérivée de la log-vraismenblance partielle par rapport à θ ;
- et $\widehat{I_n^{22}}(\widehat{\beta}, 0) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta\partial\theta^\top}\mathcal{L}_p(\beta, \theta)\right)$

▸ Retour tests HRP.

Extension : stratification

- Création de k strates relatives à k modalités d'une variable discrète ;
- Risque de base ($\alpha_0(t)$) différent selon les strates ;
- Le taux de panne pour un individu i de la strate k devient :

$$\alpha_k(t) \exp(Z_i \beta)$$

- Risques instantanés identiques pour toute les strates ;

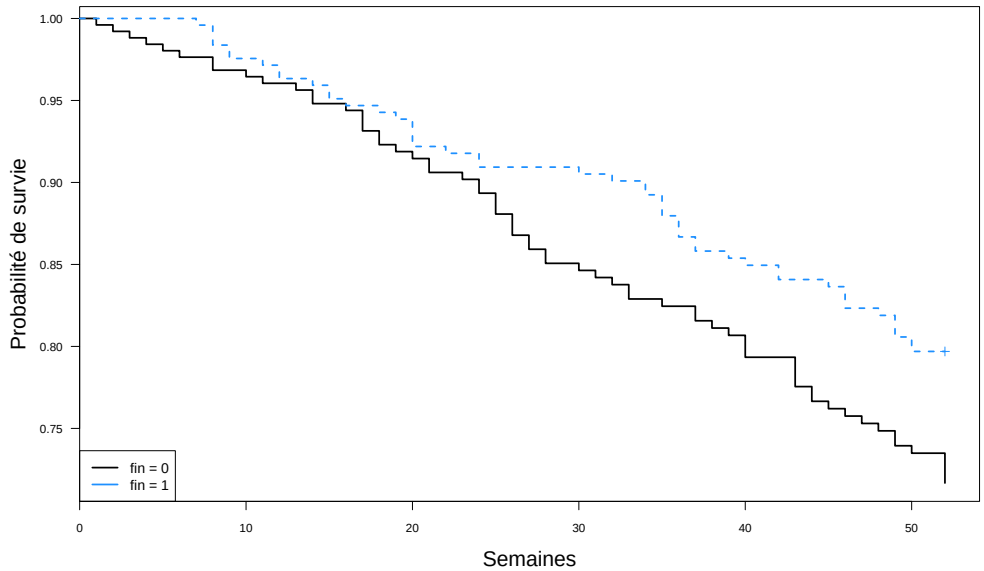

FIGURE 6: Fonctions de survie selon l'aide financière.

```
modStr <- coxph(Surv(week, arrest) ~ strata(fin) + age +
    race + wexp + mar + paro + prio, data=prison)
plot(survfit(modStr), conf.int=FALSE, ylim=range(survfit(
    modStr)$surv))
```

▸ Retour Pour aller plus loin.