

# UER Modélisation

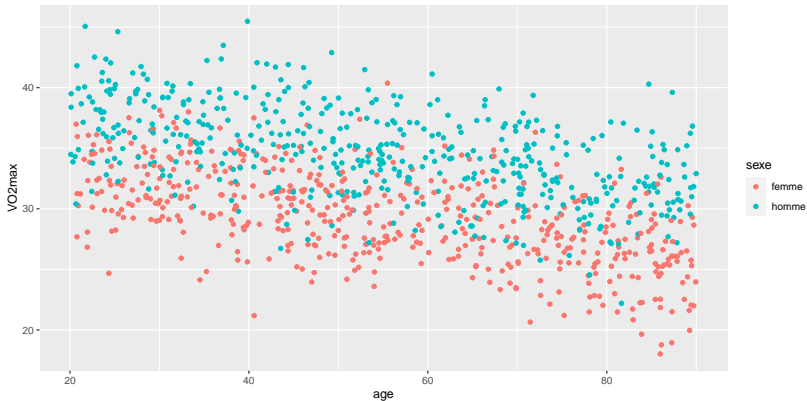
## Survie

Thomas Ferté

26/02/2022

# Rappels

# Ecrivez le modèle correspondant

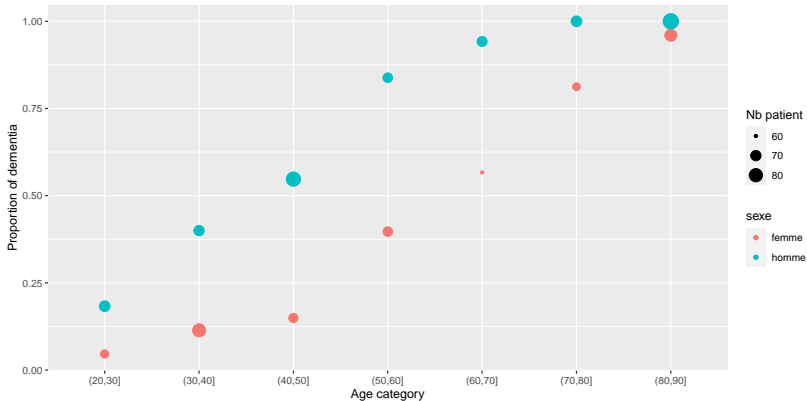


# Solution

$$VO2max_i = \beta_0 + \beta_1 age_i + \beta_2 sexe_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$$

# Ecrivez le modèle correspondant



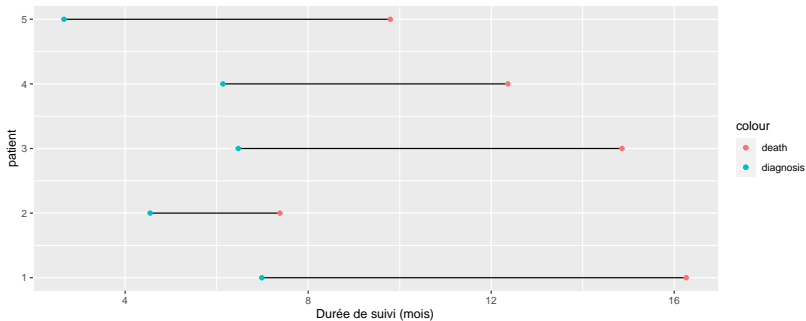
# Solution

$$\text{Logit}(P(Dementia_i = 1)) = \beta_0 + \beta_1 age_i + \beta_2 sexe_i$$

# Intuition

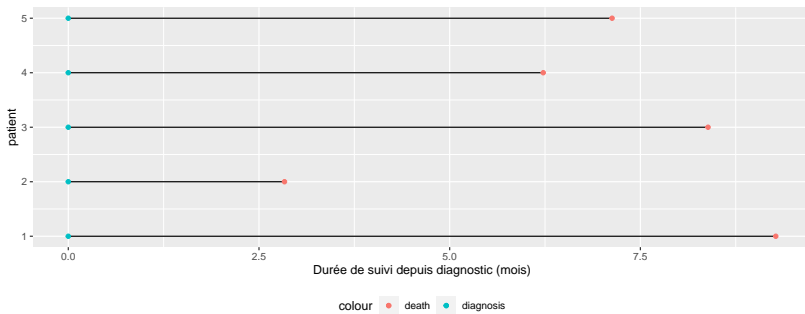
# Diagnostic et décès

On s'intéresse au lien entre le diagnostic d'un cancer et le décès d'un patient pour cela on recueille les données de plusieurs patients :

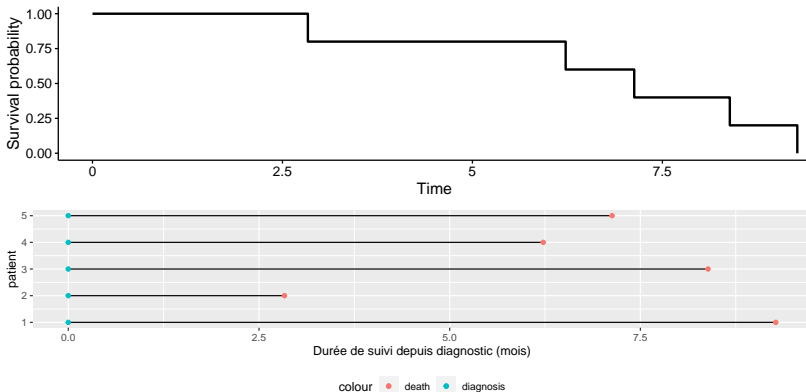




# Représentation en temps depuis diagnostic



# Représentation Kaplan Meier



Chaque marche correspond à événement et à une diminution de 20% de la probabilité de survie.

# Définitions

- *Date des dernières nouvelles* : date la plus récente où l'on a pu recueillir des informations sur un sujet.
- *Temps de participation = temps de suivi* : délai entre date d'entrée dans l'étude et la date de dernières nouvelles.
- *Date de point* : date au delà de laquelle on ne tiendra pas compte des informations sur le sujet.
- *Recul pour une étude* = date de point - date de début de l'étude
- *Recul pour un sujet* = date de point - date d'entrée dans l'étude

# Mesurer la survie

# fonction de densité

Probabilité de faire l'événement à un instant  $t$

$$f(t) = \lim_{\Delta t \rightarrow 0+} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

## fonction de répartition

Probabilité de faire l'événement avant  $t$ . Correspond au cumul de  $f(t)$  de 0 à  $t$

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \int_0^t f(u) du \end{aligned}$$

Expliquez pourquoi  $F(0) = 0$  et  $\lim_{t \rightarrow +\infty} F(t) = 1$

## fonction de survie

La fonction de survie correspond à la probabilité de ne pas avoir encore fait l'événement à l'instant  $t$

$$S(t) = P(T > t) = 1 - F(t)$$

## fonction de risque

Elle correspond à la probabilité de faire l'événement à l'instant  $t$  sachant que le patient est toujours indemne juste avant cet instant. On parle de fonction de risque instantané.

$$\begin{aligned}
 \alpha(t) &= \lim_{\Delta t \rightarrow 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\
 &= \lim_{\Delta t \rightarrow 0+} \frac{P(T \geq t | t \leq T < t + \Delta t) P(t \leq T < t + \Delta t)}{\Delta t \times P(T \geq t)} \leftarrow \text{Bayes theorem} \\
 &= \lim_{\Delta t \rightarrow 0+} \frac{1 \times P(t \leq T < t + \Delta t)}{\Delta t \times P(T \geq t)} \\
 &= \lim_{\Delta t \rightarrow 0+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \times \frac{1}{P(T \geq t)} \\
 &= \frac{f(t)}{S(t)}
 \end{aligned}$$

$$\text{Bayes theorem : } P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



## fonction de risque (2)

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u) du$$

En dérivant on obtient :  $\frac{dS(t)}{dt} = 0 - f(t) = -f(t)$

Par ailleurs

$$\begin{aligned}\alpha(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{dS(t)}{dt} \times \frac{1}{S(t)} \\ &= -\frac{d}{dt} \log(S(t))\end{aligned}$$

$$\text{NB: } \frac{d}{dx} \log(f(x)) = \frac{1}{f(x)} \times \frac{d}{dx} f(x)$$

## fonction de risque (3)

On a  $\alpha(t) = -\frac{d}{dt}\log(S(t))$  si on poursuit :

$$\alpha(t) = -\frac{d}{dt}\log(S(t))$$

$$\int_0^t \alpha(u) du = -\log(S(t))$$

$$-\int_0^t \alpha(u) du = \log(S(t))$$

$$\exp\left(-\int_0^t \alpha(u) du\right) = \exp(\log(S(t)))$$

$$\exp\left(-\int_0^t \alpha(u) du\right) = S(t)$$

## fonction de risque cumulé

Elle correspond au cumul de la fonction précédente :

$$A(t) = \int_0^t \alpha(u) du$$

## Exercice

Soit un risque instantané constant  $\alpha(t) = \lambda$

Trouvez la fonction de densité, la fonction de répartition, la fonction de survie, la fonction de risque cumulée.

Aide :

$$f(t) = S(t) \times \alpha(t)$$

$$F(t) = 1 - S(t)$$

$$S(t) = \exp\left(-\int_0^t \alpha(u) du\right)$$

$$A(t) = \int_0^t \alpha(u) du$$

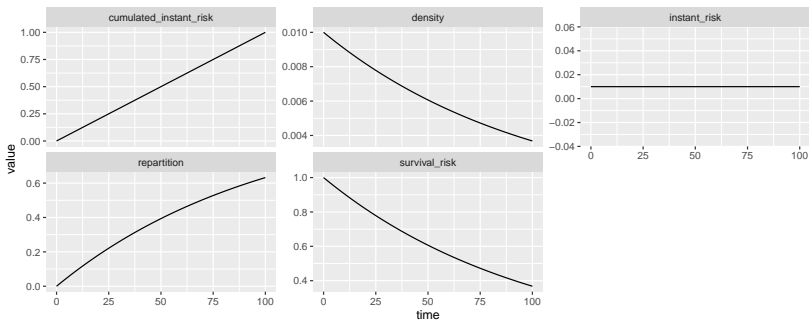
## Exercice - Solution

$$S(t) = \exp\left(-\int_0^t \alpha(u)du\right) = \exp\left(-\int_0^t \lambda\right) = \exp(-\lambda t)$$

$$A(t) = \int_0^t \alpha(u)du = \int_0^t \lambda du = \lambda t$$

$$F(t) = 1 - S(t) = 1 - \exp(-\lambda t)$$

$$f(t) = S(t) \times \alpha(t) = \exp(-\lambda t) \times \lambda$$



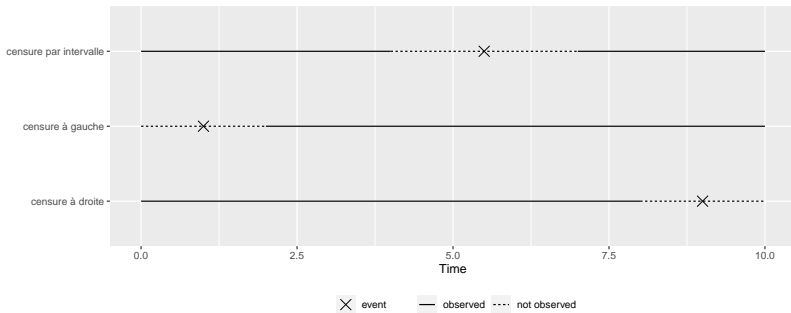
## Censure et troncature

# Censure

Lorsque l'on ne connaît pas la date précise de l'événement :

- censure à droite : l'événement survient après la fin du suivi
- censure à gauche : l'événement survient avant le début du suivi
- censure par intervalle : l'événement survient entre deux temps de suivi

# Censure en image





## Troncature

Une observation est dite tronquée si elle est conditionnelle à un autre évènement.

- Troncature à gauche : les patients de la base paquid sont recrutés parmi ceux les personnes ayant plus de 65 ans (troncature des patients de moins de 65 ans).
- Troncature à droite : très rare
- Troncature par intervalle : lorsque l'on utilise un registre, les patients ayant fait l'événement étudié par le registre avant sa mise en place ne sont pas pris en compte. Les patients répertoriés après la consultation du registre ne seront pas non plus pris en compte.

Contrairement à la censure, les patients ayant eu une troncature ne sont pas renseignés dans la base de données.

# Vraisemblance

# Vraisemblance - rappel

La vraisemblance correspond à la “probabilité” d’observer l’échantillon selon le modèle.

$$\mathcal{L} = \prod_{i=1}^n \mathcal{L}_i$$

Pour les individus  $i$  d’un échantillon de taille  $n$

# Vraisemblance - pas de censure

$$\mathcal{L}_i = f(\tilde{T}_i)$$

avec  $\tilde{T}_i$  le délai avant événement de l'individu  $i$  et  $f$  la fonction de densité de probabilité

## Vraisemblance - censure à droite

$$\mathcal{L}_i = f(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)^{1-\delta_i}$$

avec  $\tilde{T}_i$  le délai avant événement ou censure de l'individu  $i$  et  $S$  la fonction de survie.

# Exercice

On fait l'hypothèse que la survie de patients suit une loi exponentielle de paramètre  $\lambda$  avec :

$$S(t) = \exp(-\lambda t)$$

$$f(t) = \lambda \exp(-\lambda t)$$

patient	time	event
1	10	1
2	20	0
3	40	1

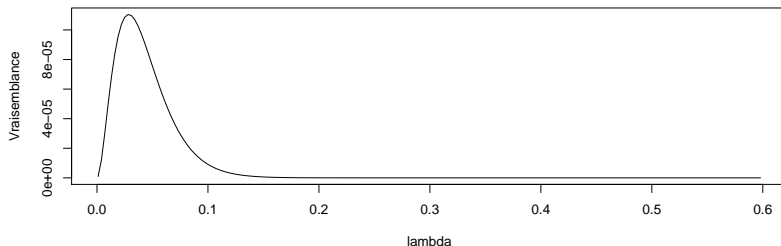
Parmi ces deux propositions, quelle valeur de  $\lambda$  vous paraît la plus probable ?

- 0.05
- 0.5

$$\text{NB : } \mathcal{L}_i = f(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)^{1-\delta_i}$$

# Solution

```
lambda = seq(0.001, 0.6, by = 0.003)
vec_vraisemblance <- sapply(lambda, function(lambda_j){
  f_ti <- lambda_j * exp(- lambda_j * dfExVraisemblance$time)
  s_ti <- exp(- lambda_j * dfExVraisemblance$time)
  vraisemblance <- prod(f_ti^dfExVraisemblance$event*s_ti^(1-dfExVraisemblance$event))
})
plot(lambda, vec_vraisemblance, ylab = "Vraisemblance", type = 'l')
```



Réponse : 0.05

## Vraisemblance - censure à droite et troncature à gauche

$$\mathcal{L}_i = \frac{f(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)^{1-\delta_i}}{S(T_{0i})}$$

avec  $\tilde{T}_i$  le délai avant événement ou censure de l'individu  $i$  et  $T_{0i}$  le délai avant troncature.

Dans la suite du cours, on ne s'intéressera qu'au cas avec censure à droite et troncature à gauche.



# Kaplan Meier

# Définition

- $d_j$  : nombre de sujets subissant l'événement au temps  $t_j$
- $n_j$  : nombre de sujets à risque au temps  $t_j$
- $\widehat{S}(t) = \prod_{j:t_j < t} \frac{n_j - d_j}{n_j}$
- D'où :  $\widehat{S}(t_{j+1}) = \widehat{S}(t_j) \times \frac{n_{j+1} - d_{j+1}}{n_{j+1}}$

# En pratique (1)

subject	time	event
1	10	1
2	15	0
3	20	1
4	22	0
5	30	0

## En pratique (2)

subject	time	event
1	10	1
2	15	0
3	20	1
4	22	0
5	30	0

time	n_j	d_j
0	5	0
10	5	1
15	4	0
20	3	1
22	2	0
30	1	0

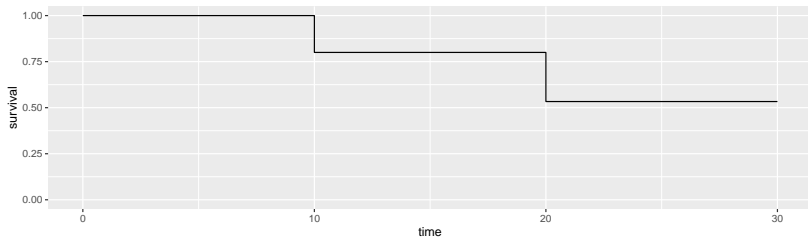
## En pratique (3)

```
dfKmAlysedStep2 <- dfKmAlysedStep1 %>%
  mutate(Prob_cond = (n_j - d_j)/n_j,
         survival = cumprod(Prob_cond))
dfKmAlysedStep2 %>% knitr::kable(booktabs = TRUE)
```

time	n_j	d_j	Prob_cond	survival
0	5	0	1.0000000	1.0000000
10	5	1	0.8000000	0.8000000
15	4	0	1.0000000	0.8000000
20	3	1	0.6666667	0.5333333
22	2	0	1.0000000	0.5333333
30	1	0	1.0000000	0.5333333

## En pratique (4)

time	n_j	d_j	Probab_cond	survival
0	5	0	1.0000000	1.0000000
10	5	1	0.8000000	0.8000000
15	4	0	1.0000000	0.8000000
20	3	1	0.6666667	0.5333333
22	2	0	1.0000000	0.5333333
30	1	0	1.0000000	0.5333333



# Exercice

Construisez l'estimateur de KM :

subject	time	event
1	2	0
2	8	1
3	10	1
4	14	0
5	16	1

# Solution

subject	time	event
1	2	0
2	8	1
3	10	1
4	14	0
5	16	1

time	n_j	d_j	Prob_cond	survival
0	5	0	1.0000000	1.00
8	4	1	0.7500000	0.75
10	3	1	0.6666667	0.50
16	1	1	0.0000000	0.00



# Intervalles de confiance

Formule de Greenwood

Formule de Rothman (++)

# Log-rank

# Intuition

Objectif : comparer deux (ou plus) courbes de survie

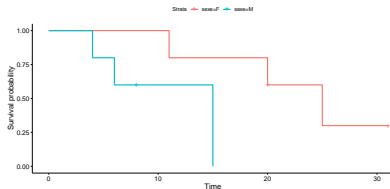
Sous  $H_0$  (pas de différence), le nombre attendu d'événement à un instant  $j$  dans chacun des groupes est similaire au nombre d'événement dans l'échantillon pondéré par le nombre d'individu dans chaque groupe.

On va comparer ce nombre attendu au nombre observé. Si la différence est trop grande alors on conclut à une différence significative.

Suit une loi du Chi-2 à  $g - 1$  où  $g$  est le nombre de groupe (e.g  $ddl = 1$  si 2 groupes)

## En pratique (1)

patient	sexe	time	died
1	M	4	1
2	M	6	1
3	M	8	0
4	F	11	1
5	M	15	1
6	M	15	1
7	F	20	1
8	F	20	0
9	F	25	1
10	F	31	0



## En pratique (2)

time	Male_at_risk	Female_at_risk	Male_deaths	Female_deaths	Deaths
4	5	5	1	0	1
6	4	5	1	0	1
11	2	5	0	1	1
15	2	4	2	0	2
20	0	4	0	1	1
25	0	2	0	1	1

## En pratique (3)

$$Expected\_death\_male = Deaths \times \frac{Male\_at\_risk}{Male\_at\_risk + Female\_at\_risk}$$

time	Male_at_risk	Female_at_risk	Male_deaths	Female_deaths	Deaths	Exp_deaths_male	Exp_deaths_female
4	5	5	1	0	1	0.5000000	0.5000000
6	4	5	1	0	1	0.4444444	0.5555556
11	2	5	0	1	1	0.2857143	0.7142857
15	2	4	2	0	2	0.6666667	1.3333333
20	0	4	0	1	1	0.0000000	1.0000000
25	0	2	0	1	1	0.0000000	1.0000000

## En pratique (4)

time	Male_at_risk	Female_at_risk	Male_deaths	Female_deaths	Deaths	Exp_deaths_male	Exp_deaths_female
4	5	5	1	0	1	0.5000000	0.5000000
6	4	5	1	0	1	0.4444444	0.5555556
11	2	5	0	1	1	0.2857143	0.7142857
15	2	4	2	0	2	0.6666667	1.3333333
20	0	4	0	1	1	0.0000000	1.0000000
25	0	2	0	1	1	0.0000000	1.0000000

$$\chi^2 = \sum_{group} \left( \frac{(\sum observed - \sum expected)^2}{\sum expected} \right)$$

$$\begin{aligned}
 \chi^2 &= \sum_{group} \left( \frac{(\sum observed - \sum expected)^2}{\sum expected} \right) \\
 &= \frac{(4 - 1.897)^2}{1.897} + \frac{(3 - 5.103)^2}{5.103} \\
 &= 3.199
 \end{aligned}$$

# Cox



# Un modèle de régression

Modèle à risque proportionnel

$$\alpha(t, Z) = \alpha_0(t) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})$$

A noter l'absence de  $\beta_0$  “remplacé” par  $\alpha_0(t)$

## Exercice

$$\alpha(t, Z) = \alpha_0(t) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})$$

Soit un modèle univarié évaluant le risque de décès en fonction du bras de traitement (0: placebo, 1: traitement).

Ecrivez le risque instantané de faire l'événement chez un patient prenant un placebo, idem chez un patient prenant le traitement.

Déduisez en le risque relatif instantané de décéder chez les patient prenant le traitement par rapport à ceux prenant le placebo.

# Solution

Modèle :  $\alpha(t, Z) = \alpha_0(t) \exp(\beta_1 \text{Traitement}_{i1})$

Patient  $j$  dans le groupe placebo :  $\alpha(t, Z_j) = \alpha_0(t) \exp(0) = \alpha_0(t)$

Patient  $k$  dans le groupe traitement :  $\alpha(t, Z_k) = \alpha_0(t) \exp(\beta_1)$

Rapport de risque :  $HR = RR = \frac{\alpha_0(t) \exp(\beta_1)}{\alpha_0(t)} = e^{\beta_1}$

# Spécification du modèle

- Choix des variables : idem régression linéaire et logistique
- Variables catégorielles : idem régression linéaire et logistique
- Modification d'effet : idem régression linéaire et logistique

# Vraisemblance partielle

Soit  $t_1, \dots, t_k$  les temps d'événements (décès) en considérant qu'il n'y a qu'un seul événement à chaque temps.

Chaque individu  $i$  a un risque instantané de faire l'événement qui dépend du temps et de ses covariables  $\alpha_i(t_i, Z_i)$ .

On peut écrire la vraisemblance telle que :

$$\begin{aligned}
 L_i(\beta) &= P(\text{individu } j \text{ fait l'événement} | \text{un des individus a fait l'événement}) \\
 &= \frac{P(\text{individu } j \text{ fait l'événement} | \text{à risque à l'instant } t_j)}{\sum P(\text{individus fassent l'événement} | \text{à risque à l'instant } t_j)} \\
 &= \frac{\alpha_0(t_j) \exp(\beta_1 X_j)}{\sum \alpha_0(t_j) \exp(\beta_1 X_l)} \\
 &= \frac{\alpha_0(t_j) \exp(\beta_1 X_j)}{\alpha_0(t_j) \sum \exp(\beta_1 X_l)} \\
 &= \frac{\exp(\beta_1 X_j)}{\sum \exp(\beta_1 X_l)}
 \end{aligned}$$

Donc pas besoin d'estimer  $\alpha_0(t)$  !!!

# Intervalle de confiance

$$IC : \exp([\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\widehat{var}(\hat{\beta}_j)}])$$

# Tests statistiques

- Test global : Rapport de vraisemblance (+++), Wald, Score
- Apport d'une variable : Wald (+++), Rapport de vraisemblance, Score
- Apport d'un ensemble de variables : Rapport de vraisemblance (+++), Wald, Score

## Tests statistiques - Wald (un seul paramètre)

Soit  $\hat{\beta}_1$  l'estimateur du coefficient  $\beta_1$  par le modèle et  $SE_{\hat{\beta}_1}$  sont erreur standard associée alors la statistique de test de wald est définie comme :

$$Wald = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

et suit une loi du Chi-2 à 1 ddl.



# Tests statistiques - Log-vraisemblance (comparaison de modèle)

Soit  $m2$  le modèle complet et  $m1$  le modèle restreint, le rapport de vraisemblance est défini comme :

$$RV = 2 \times (\loglik(m2) - \loglik(m1))$$

Dans ce cas  $RV$  suit une loi du Ch-2 avec un ddl égal à la différence du nombre de paramètres ( $\beta$ ) entre les deux modèles.

## Tests statistiques - Score (un seul paramètre)

Soit  $\beta$  le paramètre du modèle de cox à tester. Soit  $U(\beta)$  la dérivée première de la vraisemblance du modèle selon ce paramètre et  $I(\beta)$  l'opposée de l'espérance de la dérivée seconde de la vraisemblance de ce paramètre. Alors la statistique du score est définie telle que :

$$Score = \frac{U(\beta)^2}{I(\beta)}$$

et suit une loi du Chi-2 à 1 ddl

# Hypothèses du modèle

# Introduire le temps dans un modèle de survie

Fin

Questions ?