Durée de vie des Motorettes
(analyse de survie)

GMMA 106

Cas d'étude

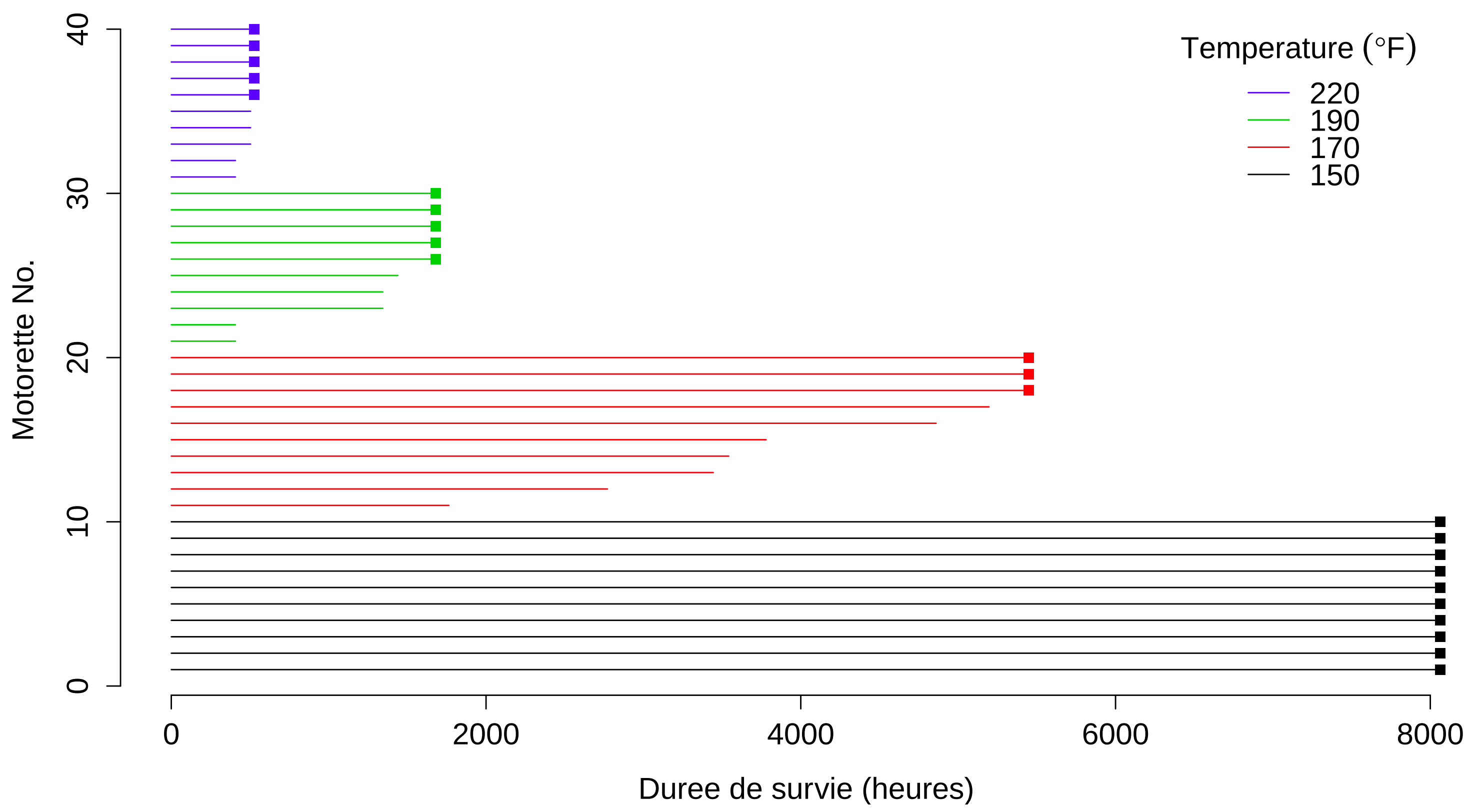

Figure 1: *Durées de fonctionnement avant défaillance de 40 motorettes.*

- Quels commentaires souhaitez vous faire ?

Les données

```
> library(SMPracticals); data(motorette)
> motorette
     x cens    y
1  150    0 8064
2  150    0 8064
3  150    0 8064
.
.
38 220    0  528
39 220    0  528
40 220    0  528
```

x Température °F

cens Variable indicatrice de censure (à droite), 1: non censurée / 0: censurée

y Durée de survie en heure

Objectifs et éléments parcourus

- Notre objectif est de proposer un modèle statistique pour modéliser la durée de fonctionnement avant défaillance de motorettes en fonction de la température.
- Le but initial de cette expérience était de caractériser le comportement à $130°$F—mais le faire à cette température aurait été trop long/coûteux !
- Ceci nous permettra de croiser les objets suivants :
  - censure à droite, gauche, par intervalle, de type I et de type II, notations conventionnelles
  - fonction de survie, taux de panne, taux de panne cumulé
  - Kaplan–Meier
  - Vraisemblance pour des données censurées

1. Analyse de survie et données censurées

- L'analyse de survie consiste à analyser la durée avant qu'un événement précis se produise.
- Voici quelques exemples
  - Durée avant la résurgence d'une tumeur
  - Durée avant qu'un composant électronique casse
  - Durée avant qu'un étudiant s'endorme pendant ce cours...
- L'analyse de survie est spécifique car les données sont très particulières
  - positives et fortement asymétriques
  - l'intérêt porte sur une durée éloignée plutôt qu'un comportement moyen
  - présence de censure

Définition 1. Une observation est dite censurée lorsque cette dernière est partiellement connue. Par exemple l'observation est plus grande/petite qu'une valeur seuil; ou encore l'observation est comprise dans l'intervalle $[a, b]$.

Dans un contexte médical, la censure peut apparaître lorsque

- le patient quitte l'étude
- le patient ne montre pas de symptômes/progrès avant la fin de l'étude
- le fichier du patient est perdu...

Les différents types de censure

- Les observations peuvent présenter différents types de censure

  gauche l'observation est plus petite qu'une valeur mais de combien ?

  intervalle l'observation est tombée quelque part dans un intervalle donnée mais où ?

  droite l'observation est plus grande qu'une valeur mais de combien ?

- L'expérience elle-même peut engendrer cette censure

  type 1 L'expérience à une durée prédéterminée provoquant éventuellement une censure à droite

  type 2 L'expérience s'arrête lorsqu'un nombre prédéterminé d'unités/patients ont faillies.

Représentation des observations censurées

- Il existe une convention afin d'écrire le type de censure des observations. Ainsi

  4 indique que l'observation 4 est non censurée
  4+ indique que l'observation 4 est censurée à droite
  4- indique que l'observation 4 est censurée à gauche
  [5-10] indique une censure par intervalle

- Ainsi nos données peuvent s'écrire dans le tableau suivant

Table 1: *Données de survie motorette.*

°F	Durée de fonctionnement avant défaillance (heures)									
150	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+
170	1764	2772	3444	3542	3780	4860	5196	5448+	5448+	5448+
190	408	408	1344	1344	1440	1680+	1680+	1680+	1680+	1680+
220	408	408	504	504	504	528+	528+	528+	528+	528+

Soient $T_1, \ldots, T_n$ n durées de survie et C_i le temps de censure (à droite) pour le i-ème sujet. La réponse observée est alors donnée par

$$T_i^* = \min(T_i, C_i),$$

ou sous une autre forme

$$T_i^* = T_i 1_{\{T_i \leq C_i\}} + C_i 1_{\{T_i > C_i\}}.$$

Définition 2. Soit T de densité f et f.d.r. F. On appelle

a) fonction de survie (survival function)

$$S(t) = \Pr[T > t] = 1 - F(t) \in [0, 1]$$

b) taux de panne (hazard function)

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[t < T < t + \Delta t \mid T > t]}{\Delta t} = \frac{f(t)}{S(t)} > 0$$

c) taux de panne cumulé (cumulative hazard function)

$$H(t) = \int_0^t h(u) \mathrm{d}u > 0.$$

Liens entre ces fonctions

Remarque. La connaissance d'une seule de ces fonctions permet de connaître les autres. En effet

$$\begin{aligned} h(t) &= -\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t) \\ H(t) &= -\ln S(t) \\ S(t) &= \exp\{-H(t)\}. \end{aligned}$$

□ Ainsi on pourra poser un modèle statistique paramétrique en faisant un choix particulier sur l'une ou l'autre des fonctions.

- Puisque $S(t) = \Pr[T > t]$, une estimation non paramétrique est
$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} 1_{\{t \leq T_i\}}$$
- Mais cet estimateur n'est pas pertinent en présence de censure
- Il est plus adapté d'utiliser l'estimateur de Kaplan–Meier
- D'autres estimateurs existent "Nelson–Aalen" ou "life table" mais nous ne les verrons pas

Exercice 1. Soient $0 = t_0 < t_1 < t_2 < \cdots < t_n$ les durée de survie ordonnées. Montrez que pour $k \in \{1, \ldots, n\}$,

$$S(t_k) = \prod_{j=1}^{k} \Pr[T > t_j \mid T > t_{j-1}].$$

Exercice 1. Soient $0 = t_0 < t_1 < t_2 < \cdots < t_n$ les durée de survie ordonnées. Montrez que pour $k \in \{1, \ldots, n\}$,

$$S(t_k) = \prod_{j=1}^{k} \Pr[T > t_j \mid T > t_{j-1}].$$

- ☐ Les quantités $\Pr[T > t_j \mid T > t_{j-1}]$ sont estimées par

$$1 - d_j/n_j,$$

où n_j est le nombre de sujets encore à risque/en vie pendant $[t_{j-1}, t_j)$ et d_j le nombre de défaillances/morts au temps t_j.

Exemple

Soit les données de survie suivantes $6, 8+, 12, 3, 21, 12$. D'où

j	t_j	n_j	d_j	$1 - d_j/n_j$	$\hat{S}(t_j)$
0	0	6	0	1	1
1	3	6	1	$1 - 1/6$	$5/6$
2	6	5	1	$1 - 1/5$	$4/5 \times 5/6$
3	8+	4	0	1	$4/5 \times 5/6$
4	12	3	2	$1 - 2/3$	$2/3 \times 4/5 \times 5/6$
5	21	1	1	0	$0 \times 2/3 \times 4/5 \times 5/6$

Remarque. Notez comment agit l'observation censurée. La procédure suppose qu'il y a eu défaillance/mort pendant le laps de temps $(8, 12)$ mais n'est pas comptabilisée dans les d_j.

Exercice 2. Trouvez l'estimateur de Kaplan–Meier pour les données motorette à $190°$F et le code `R` associé.

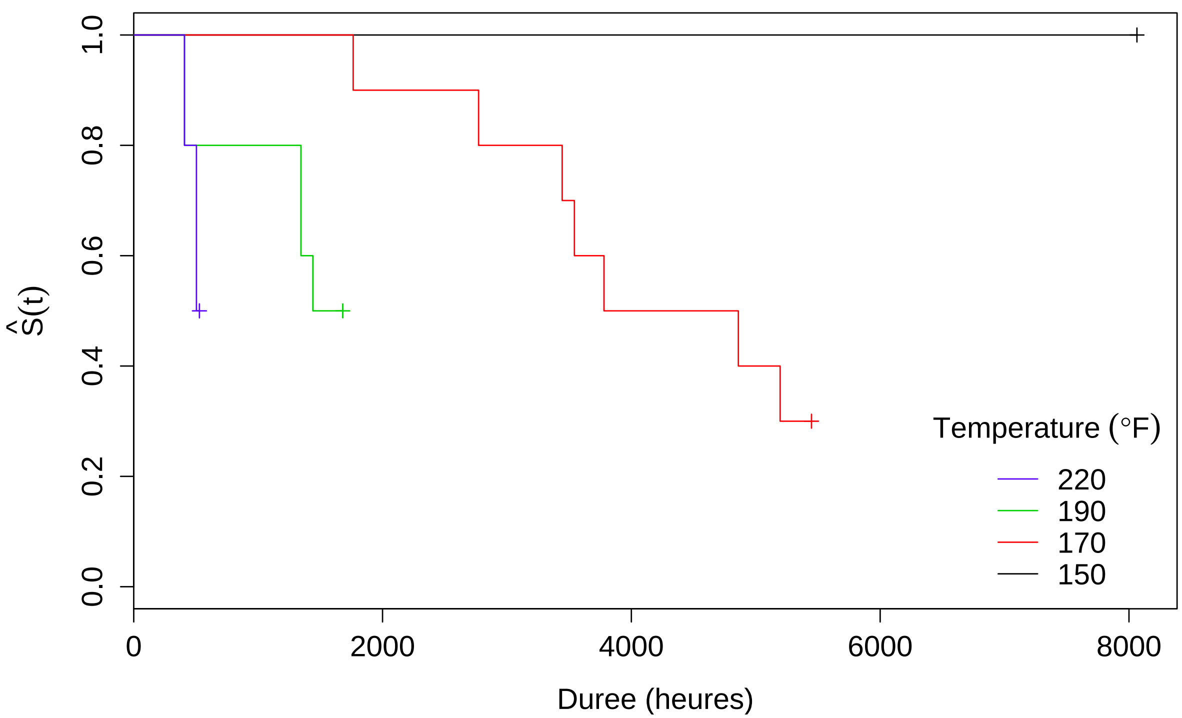

Figure 2: *Estimateur de Kaplan–Meier pour la fonction de survie des données motorette.*

- Le symbole + code la présence d'observations censurées.
- Pourquoi une estimation pour chaque température ?

Hypothèses pour Kaplan–Meier

Il faut garder en mémoire que l'estimateur de Kaplan–Meier repose sur deux hypothèses importantes :

censure non informative Suppose l'indépendance entre T_i et C_i dans la représentation $T_i^* = \min(T_i, C_i)$. Si elle ne pas vérifiée, estimateur biaisé.

homogénéité Chaque T_i provient de la même loi. Si ce n'est pas le cas, risque de mauvaise interprétation/non sens...

2. Analyses de survie paramétriques

$$f(t) = \lambda \exp(-\lambda t), \qquad S(t) = \exp(-\lambda t)$$
$$h(t) = \lambda, \qquad H(t) = \lambda t$$

Exercice 3. Montrez qu'un modèle exponentiel est sans mémoire, i.e.,

$$\Pr[T > t + \Delta \mid T > t] = \Pr[T > \Delta], \qquad \Delta, t > 0.$$

$$f(t) = \lambda \exp(-\lambda t), \qquad S(t) = \exp(-\lambda t)$$
$$h(t) = \lambda, \qquad H(t) = \lambda t$$

Exercice 3. Montrez qu'un modèle exponentiel est sans mémoire, i.e.,

$$\Pr[T > t + \Delta \mid T > t] = \Pr[T > \Delta], \qquad \Delta, t > 0.$$

- Le taux de panne h est constant—souvent peu réaliste

Modèle de Weibull $(\lambda > 0, \kappa > 0)$

$$f(t) = \kappa\lambda^{-\kappa}t^{\kappa-1}\exp\{-(t/\lambda)^{\kappa}\}, \qquad S(t) = \exp\{-(t/\lambda)^{\kappa}\}$$
$$h(t) = \kappa\lambda^{-\kappa}t^{\kappa-1}, \qquad H(t) = \lambda^{-\kappa}t^{\kappa}$$

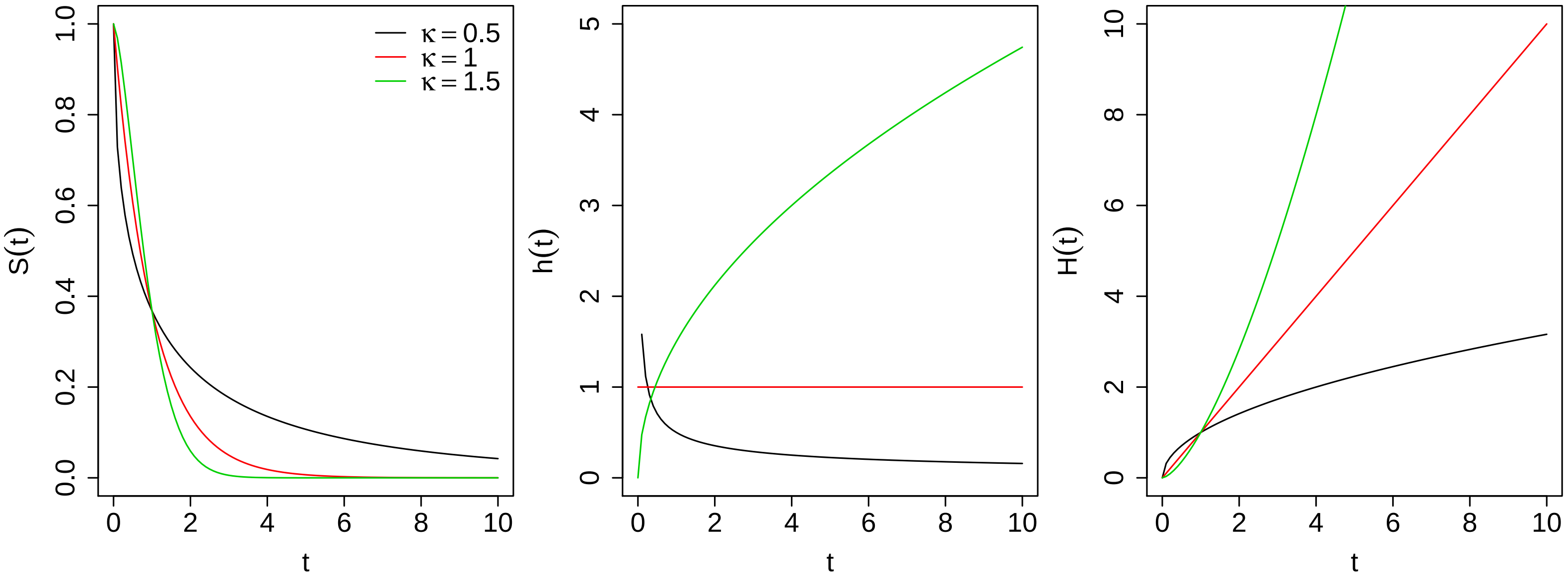

Figure 3: *Graphes de la fonction de survie $S(t)$, du taux de panne $h(t)$ et du taux de panne cumulé $H(t)$ pour le modèle de Weibull avec $\lambda = 1$ et $\kappa = 0.5, 1, 1.5$.*

Estimation d'un modèle paramétrique de survie

- Nous allons utiliser l'estimateur du maximum de vraisemblance
- Attention à la présence de données censurées. La contribution de la i-ème observation t_i à la vraisemblance est
  - $f(t_i; \theta)$ si t_i n'est pas censurée
  - $\Pr[T > t_i] = 1 - F(t_i)$ si t_i est censurée à droite
  - $\Pr[T < t_i] = F(t_i)$ si t_i est censurée à gauche
  - $\Pr[a < T < b] = F(b) - F(a)$ si t_i est censuré par l'intervalle $[a_i, b_i]$

Estimation d'un modèle paramétrique de survie

- Nous allons utiliser l'estimateur du maximum de vraisemblance
- Attention à la présence de données censurées. La contribution de la i-ème observation t_i à la vraisemblance est
  - $f(t_i;\theta)$ si t_i n'est pas censurée
  - $\Pr[T > t_i] = 1 - F(t_i)$ si t_i est censurée à droite
  - $\Pr[T < t_i] = F(t_i)$ si t_i est censurée à gauche
  - $\Pr[a < T < b] = F(b) - F(a)$ si t_i est censuré par l'intervalle $[a_i, b_i]$
- La vraisemblance est alors

$$L(\theta) = \prod_{i \in \mathcal{NC}} f(t_i;\theta) \prod_{i \in \mathcal{D}} \{1 - F(t_i)\} \prod_{i \in \mathcal{G}} F(t_i) \prod_{i \in \mathcal{I}} \{F(b_i) - F(a_i)\},$$

où $\mathcal{NC}$ désigne l'ensemble des observations non censurées, $\mathcal{D}$ celles à droite, $\mathcal{G}$ à gauche et $\mathcal{I}$ par intervalle.

3. Ce que nous ne verrons pas

Pour les personnes intéressées

Nous avons à peine survolé l'analyse de survie. Les personnes intéressées pourrant regarder les thèmes suivants :

- Comparaison de fonctions de survie
- Modèles de Cox, modèles de fragilité (frailty models)
- Modèle log-logistique

4. Les mains dans le cambouis

Modèle

°F	Durée de fonctionnement avant rupture (heures)									
150	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+
170	1764	2772	3444	3542	3780	4860	5196	5448+	5448+	5448+
190	408	408	1344	1344	1440	1680+	1680+	1680+	1680+	1680+
220	408	408	504	504	504	528+	528+	528+	528+	528+

Nous allons considérer le modèle de Weibull suivant

$$\Pr[T_{i,j} \leq t; x_i] = 1 - \exp\left\{-(y/\theta_i)^\gamma\right\}, \qquad \theta_i = \exp(\beta_0 + \beta_1 x_i),\ \gamma > 0,$$

pour $i = 1, \ldots, 4$, $j = 1, \ldots, 10$ et où les durées à défaillances seront prises en centaines d'heures et la covariable x_i est $\ln$(température$/100$).

Ce que j'aimerais que vous fassiez

- ☐ Écrire la vraisemblance pour ce modèle
- ☐ Écrire une fonction R calculant l'estimateur du maximum de vraisemblance et sa variance.
- ☐ Commentez vos résultats et faire de jolies représentations graphiques