EXAMEN FINAL

Modèle de Durée– 2023-2024
Pierre-O Goffard

Instructions: On éteint et on range son téléphone.

- La calculatrice et les appareils éléctroniques ne sont pas autorisés.
- Vous devez justifier vos réponses de manière claire et concise.
- Vous devez écrire de la manière la plus lisible possible. Souligner ou encadrer votre réponse finale.
- Document autorisé: Une feuille manuscrite recto-verso

Question:	1	2	3	Total
Points:	4	8	8	20
Score:				

1. (4 points) Nous souhaitons calibrer sur nos données un modèle exponentiel Exp(β) de densité

$$f(t) = \frac{e^{-t/\beta}}{\beta}\mathbb{I}_{(0,\infty)}(t),$$

avec $\beta > 0$.

Nous disposons d'un échantillon de taille n. Nos données sont à la fois tronquées à gauche, avec un niveau de troncature $c > 0$ pour toutes les observations et censurées à droites, avec un niveau de censure $c_1, \dots, c_n$, différent pour chaque observation tels que $c_i > c$ pour tout $i = 1, \dots, n$.

Donner l'estimateur du maximum de vraisemblance pour le paramètre β. Il faut rappeler les notations pour les données censurées et détailler les étapes de calcul menant à l'expression de l'estimateur.

Solution: Les données sont notée

$$\mathcal{D} = (x_i, \delta_i)_{i=1,\dots,n} = (t_i \wedge c_i, \mathbb{I}_{t_i \leqslant c_i})_{i=1,\dots,n},$$

où t_i sont les observations non censurées. Ces observations non-censurées sont tronquées à gauche au niveau c, leur densité est donnée par

$$f_{(c,\infty)}(t) = \frac{e^{-(t-c)/\beta}}{\beta}\mathbb{I}_{(c,\infty)}(t),$$

leur fonction de survie par

$$S_{(c,\infty)}(t) = \begin{cases} 1, & sit \leqslant c, \\ e^{-(t-c)/\beta}, & t > c. \end{cases}$$

et leur fonction de hasard est donnée par

$$h_{(c,\infty)}(t) = \frac{1}{\beta}\mathbb{I}_{(c,\infty)}(t).$$

La log-vraisemblance des observation est donnée par

$$l(\mathcal{D};\beta) = \frac{1}{\beta}\sum \delta_i - \frac{1}{\beta^2}\sum(x_i - c)$$

L'équation

$$\frac{\partial}{\partial\beta} l(\mathcal{D};\beta) = 0,$$

équivaut à

$$\hat{\beta} = \frac{\sum(x_i - c)}{\sum \delta_i}.$$

On peut vérifier que

$$\frac{\partial^2}{\partial\beta^2} l(\mathcal{D};\beta)\Big|_{\beta=\hat{\beta}} < 0$$

2. Nous avons des données de mortalité pour la Belgique pendant l'année 2000 pour des personnes agées de 30 à 105 ans, voir le Tableau 1.

Age	Year	E_x^c	D_x
30	2000	146944	119
31	2000	147252	135
32	2000	148805	132
33	2000	152792	160
34	2000	157925	161
35	2000	163324	155

Table 1: Nombre de décès et expositions centrales pour la Belgique pendant l'année 2000.

(a) (1 point) Nous allons estimer les taux de mortalité via le modèle de Poisson. Rappeler les hypothèses du modèle et l'expression de l'estimateur des taux de mortalités.

Solution: Dans le cadre du modèle de Poisson, on suppose que

$$D_x \sim \text{Pois}()\mu_x \cdot E_x^c),$$

où D_x est le nombre de décès parmi les individus d'âge x, E_x^c est l'exposition centrale pour les individus d'âge x et μ_x est le taux de mortalité pour les individus d'âge x. Les taux de mortalités sont estimés par

$$\hat{\mu}_x = \frac{D_x}{E_x^c}.$$

(b) (1 point) La Figure 1 montre les taux de mortalités estimés en fonction de l'âge. Nous

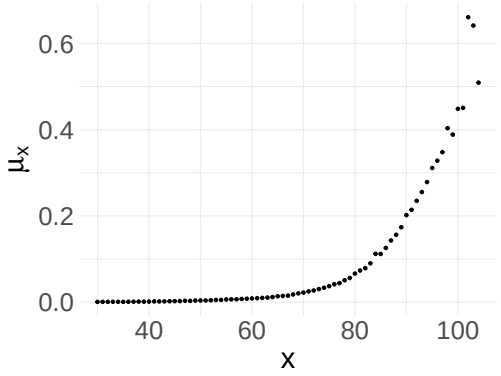

Figure 1: Taux de mortalité en fonction de l'âge.

allons ajuster le modèle suivant:

$$\mu_x = \left(\frac{\alpha}{\beta}\right) \cdot \left(\frac{x}{\beta}\right)^{\alpha-1}, \tag{1}$$

sur les taux estimés de la Figure 1. De quel type de procédure s'agit-il? Quel est l'intérêt d'une telle procédure?

Solution: Il s'agit d'une procédure de lissage paramétrique. Elle permet de remplacer les taux de mortalités aux grands âge peu fiable du fait de la faible exposition.

(c) (1 point) La Figure 2 montre le logarithme des taux de mortalités en fonction du log de l'âge.
En quoi ce graphique conforte le choix du modèle de la question b)?

Solution: En prenant le log dans l'équation (1), il vient

$$\log(\mu_x) = \log(\alpha) - \alpha\log(\beta) + (\alpha - 1)\log(x).$$

Notre modèle de lissage suppose donc un lien linéaire entre le log des taux de mortalité et le log de l'âge. On constate peu ou proue ce lien sur la Figure 2.

(d) (1 point) Les données de mortalité du Tableau 1 sont stockées dans un data frame R, appelé `df_Bel`. Nous faisons tourné le programme suivant:

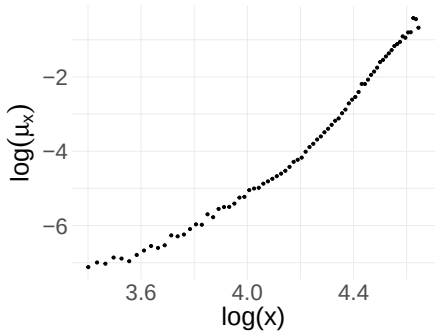

Figure 2: Log des taux de mortalité en fonction du log de l'âge.

```
df_Bel_V1 <- df_Bel %>%
mutate(mux = Dx / ExC, logx = log(Age), logmux = log(mux))

res <- lm(logmux ~ logx, data = df_Bel_V1)
```

Le contenu de l'objet `res` est donné sur la Figure 3.

```
Call:
lm(formula = logmux ~ logx, data = df_Bel_V1)

Coefficients:
(Intercept)         logx
    -27.455        5.671
```

Figure 3: Contenu de l'objet `res`.

Peut-on déduire la valeur de α et β de la sortie R de la Figure 3? Expliquer.

Solution: En notant $a = -27.455$ et $b = 5.671$, on identifie

$$\begin{cases} a = \log(\alpha) - \alpha \log(\beta) \\ b = \alpha - 1 \end{cases}$$

ce qui équivaut à

$$\begin{cases} \alpha = b + 1 \\ \beta = \exp\left(\frac{\log(b+1) - a}{b+1}\right) \end{cases}$$

(e) (2 points) Les taux de mortalités issus du modèle de la question b) sont donnés sur la Figure 4.

Que pensez vous de cet ajustement ? Comment pourrait-on l'améliorer? On pourra s'appuyer sur les Figures 2 et 4 pour répondre. Pour information $\exp(4) \approx 54.6$.

Solution: Cet ajustement est assez peu satisfaisant car il conduit à une sur-estimation des taux de mortalités aux jeunes âges et une sur-estimation au âges plus avancés. On

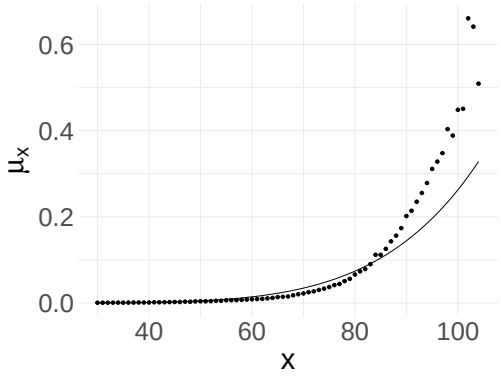

Figure 4: Taux de mortalité estimé (points) et ajustés avec le modèle de la question b) (trait plein).

remarque sur la figure 2 une cassure pour $log(x) = 4$ avec potentiellement deux tendances linéaires distinctes pour les plages $x \in (30, e^4)$ et $x \in (e^4, 105)$. Une solution serait d'ajuster le modèle séparément sur ces deux plages ou bien seulement remplacer les taux de mortalités à partir de $x > e^4$ par ceux du modèle calibré sur la plage $x \in (e^4, 105)$.

(f) (2 points) En admettant que l'on ait été en mesure de déterminer α et β, nous les stockons dans les variables `alpha` et `beta`. Cela nous permet de calculer les taux de mortalités ajustés par le modèle de la question b). On fait tourner le programme R suivant:

```
df_Bel_smooth <- df_Bel_V1 %>%
mutate( mux_smooth = ( alpha / beta) * (Age / beta)^(alpha - 1))

res_binom_test <- binom.test(
sum(df_Bel_smooth$mux > df_Bel_smooth$mux_smooth),
nrow(df_Bel_smooth),
p=0.5)
```

Le contenu de l'objet `res_binom_test` est donné par la Figure 5.

```
        Exact binomial test

data:  sum(df_Bel_smooth$mux > df_Bel_smooth$mux_smooth) and nrow(df_Bel_smooth)
number of successes = 33, number of trials = 75, p-value = 0.3557
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3254672 0.5594178
sample estimates:
probability of success
                  0.44
```

Figure 5: contenu de l'objet `res_binom_test`.

De quelle procédure s'agit-il et quel est son intérêt? Quelle interprétation faites-vous du résultat?

Solution: Il s'agit d'un test des signe permettant de valider la méthode de lissage en vérifiant que la procédure de lissage ne sous ou sur-estime pas systématiquement les taux de mortalité brut. Nous observon que la p-valeur ne permet de rejeter H_0 suivant laquelle la probabilité de sur ou de sous-estimation est égale à 0.5. Le test des signe est un succès, on pouvait s'y attendre puisque qu'on sur estime pour les jeunes âges et sous-estime pour les grnds âges. Un test des run pourrait être effectué en complément, il conduirait à indiqué la présence de sur-estimation ou de sous-estimation successive qui démontrerait que la méthode de lissage n'est pas optimale.

3. Les données contiennent des mesures effectuées sur des patients atteints de mélanome. Chaque patient a subi une ablation de la tumeur par chirurgie au Département de Chirurgie Plastique de l'Hôpital Universitaire d'Odense, au Danemark, pendant la période de 1962 à 1977. Un extrait des données est fourni par le Tableau 2.

time	status	sex	age	year	thickness	ulcer
10	3	1	76	1972	6.76	1
30	3	1	56	1968	0.65	0
35	2	1	41	1977	1.34	0
99	3	0	71	1968	2.90	0
185	1	1	52	1965	12.08	1
204	1	1	28	1971	4.84	1

Table 2: Extrait des données de patients ayant été opéré pour un mélanome

Voici une courte description des variables:

- `time`: Temps de survie depuis l'opération (nombre de jours)
- `status`: Variable catégorielle indiquant si le patient est mort à cause du mélanome (1), a survécu jusqu' à la fin de l'étude (2), est mort d'une cause autre que le mélanome (3).
- `sex`: Genre du patient (0 = Femme et 1 = Homme)
- `age`: Age du patient en année
- `year`: Année de l'opération
- `thickness`: Epaisseur de la tumeur en millimètres
- `ulcer`: Présence (1) ou Absence (0) d'un ulcère sur la tumeur

Nous allons considérer la survie des patients sans distinguer la cause du décès entrainant le recodage suivant de la variable `status`:

```
melanoma <- melanoma %>%
  mutate(status_os = if_else(status == 2, 0, 1))
```

(a) (2 points) Nous nous intéressons d'abord à l'impact variable `ulcer` sur la survie des patients. Nous commençons par faire tourner le code suivant:

```
fit <- survfit(Surv(time, status_os) ~ ulcer, data = melanoma)
ggsurvplot(fit, pval = FALSE, conf.int = TRUE, linetype = "strata",
 ggtheme = theme_minimal())
```

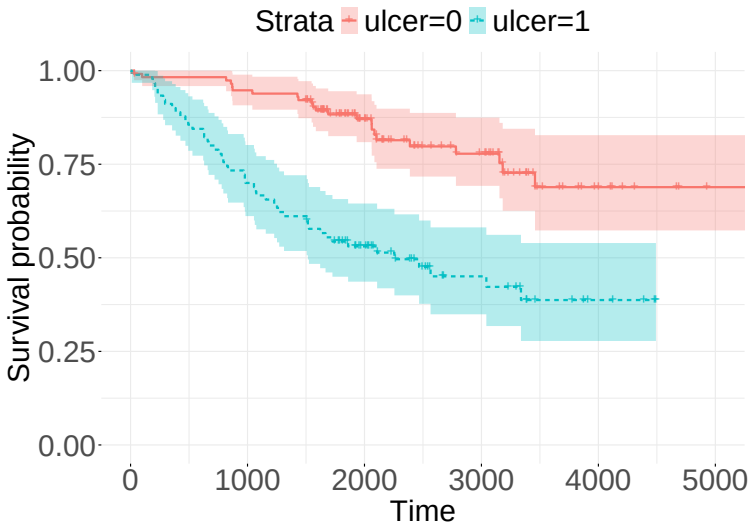

Figure 6: Sortie du programme R.

Le résultat est donné sur la Figure 6. Que représente ce graphique? Quels commentaires pouvez vous faire?

Solution: Le graphique montre les estimateurs de Kaplan-Meier des fonctions de survie des sous-populations dicriminées par la variable `ulcer`. Les courbes de survies sont bien distinctes. On constate que la présence d'un ulcère augmente la probabilité de décès, la probabilté de survie étant inférieur lorsque `ulcer` $= 1$.

(b) (2 points) Nous poursuivons notre analyse avec le programme suivant:

```
surv_diff <- survdiff(Surv(time, status_os) ~ ulcer, data = melanoma)
surv_diff
```

Le résultat est donné par la Figure 7.

```
Call:
survdiff(formula = Surv(time, status_os) ~ ulcer, data = melanoma)

          N Observed Expected (O-E)^2/E (O-E)^2/V
ulcer=0 115       23     44.5      10.4      27.9
ulcer=1  90       48     26.5      17.3      27.9

 Chisq= 27.9  on 1 degrees of freedom, p= 1e-07
```

Figure 7: Sortie du programme R.

Quelle procédure a-t-on utilisée? Quelle est votre interprétation des résultats?

Solution: Nous effectuons ici un test du log rang permettant de tester l'hypothèse H_0 suivant laquelle les deux courbes de survie sont identiques. La p-valeur conduit à rejeter H_0 et à conclure de l'impact significative de la variable `ulcer` sur la survie des patinets atteints d'un mélanome.

(c) (4 points) Nous concluons notre étude avec le programme suivant:

```
res.cox <- coxph(
Surv(time, status_os) ~ ulcer + year + age + sex + thickness,
data = melanoma)
```

Le résultat est donné par la Figure 8.

```
Call:
coxph(formula = Surv(time, status_os) ~ ulcer + year + age +
    sex + thickness, data = melanoma)

               coef exp(coef)  se(coef)      z        p
ulcer      0.976403  2.654888  0.267794  3.646 0.000266
year      -0.088291  0.915495  0.055140 -1.601 0.109328
age        0.025218  1.025539  0.007902  3.191 0.001416
sex        0.427216  1.532983  0.239618  1.783 0.074602
thickness  0.092146  1.096525  0.034882  2.642 0.008251

Likelihood ratio test=50.4  on 5 df, p=1.146e-09
n= 205, number of events= 71
```

Figure 8: Sortie du programme R.

Quelle procédure a-t-on utilisée (Rappeler les hypothèses du modèle et l'objectif du modèle)? Quelle est votre interprétation des résultats?

Solution: Nous calibrons ici un modèle de hasard proportionnelle qui spécifie la fonction de hasard conditionnellement aux covariables x par

$$h(t|x) = h_0(t)e^{\beta x},$$

où h_0 désigne la fonction de hasard de base et le vecteur β représente les paramètres du modèle qui caractérise l'impact de chacune des variable sur la fonction de hasard et donc le risque de décès. L'inspection des p-valeur indique que les variable year et sex n'ont pas un impact significatif sur la survie. On observe que la probabilité de décès augmente avec l'âge et la taille de la tumeur. Cette probabilité qugmente également lorsqu'un ulcère est présent. En efet l'exponentielle de ce coefficient est supérieur à 1 ce qui contribue à augmenter le risque de base.