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	6	5	15
Score:				

1. Le temps avant la défailance d'un composant électronique est modélisé via la variable aléatoire $T \sim \text{Lomax}(\alpha)$ dont la densité est donnée par

$$f(t) = \alpha(t+1)^{-\alpha-1}\mathbb{I}_{(0,\infty)}(t).$$

   (a) (1 point) Donner la fonction de survie et la fonction de hasard de T.

   Solution:

   $$S(t) = (1+t)^{-\alpha} \text{ (0.5 point)}, \text{ et } h(t) = \alpha(1+t)^{-1} \text{ (0.5 point)}$$

   (b) (1 point) Donner l'expression de la vraisenblance d'un échantillon de n observations de T, censurées à droite, de censure non informative. On prendra soin de rappeler les notations et leur signification.

   Solution: Voir le cours

   (c) (1 point) Donner l'expression de l'estimateur du maximum de vraisemblance de α en présence de n observations censurées à droite.

Solution:

$$\widehat{\alpha} = \frac{\sum_{k=1}^{n} \delta_k}{\sum_{k=1}^{n} \ln(1+x_k)}$$

(d) (1 point) Donner un intervalle de confiance pour α

Solution: L'estimateur du maximum de vraisemblance est asymptotiquement normal avec

$$\widehat{\alpha} \sim \text{Normal}(\alpha, I_n(\alpha)^{-1})$$

où

$$I_n(\alpha) = -\frac{\partial}{\partial \alpha} l(\mathcal{D}, \alpha)|_{\alpha=\widehat{\alpha}}$$

puis

$$\alpha \in \left[\widehat{\alpha} \pm q_{1-\alpha/2} \frac{\sqrt{\sum \delta_i}}{\sum \log(1+x_i)} \right],$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi normale, avec $\alpha = 0.05$ par exemple.

2. Nous disposons de données de nombre de décès D_x et de nombre d'individu E_x^0 d'âge $x \in [0, 100]$ en Islande au début de l'année 2010.

(a) (1 point) Nous décidons d'étudier la mortalité au moyen d'un modèle binomial. Rappeler les notations, hypothèses et estimateur associés à un tel modèle.

Solution: Dans le modèle binomial, on suppose que le nombre de décès suuit une loi binomial tel que

$$D_x \sim \text{Binom}(E_x^0, q_x),$$

L'estimateur de la probabilité de décès est donné par

$$\widehat{q}_x = \frac{D_x}{E_x^0},$$

en supposant ici que le nombre d'individu d'âge x correspond à l'exposition initiale.

(b) (1 point) Le graphique des probabilités de décès est donnée sur la Figure 1. A quoi correspondent les lignes en pointillé? Comment les obtient-on?

Solution:

- Les lignes en pointillé correspondent à l'intervalle de confiance de la probabilité de décès (0.5 points)
- L'estimateur de la probabilité de décès est asymptotiquement normal (lorsque $E_x^0 \to \infty$) avec

$$\widehat{q}_x \sim \text{Normal}\left(q_x, \frac{\widehat{q}_x(1-\widehat{q}_x)}{E_x^0} \right),$$

on en déduit que

$$q_x \in \left[q_x \pm q_{1-\alpha/2} \sqrt{\frac{\widehat{q}_x(1-\widehat{q}_x)}{E_x^0}} \right],$$

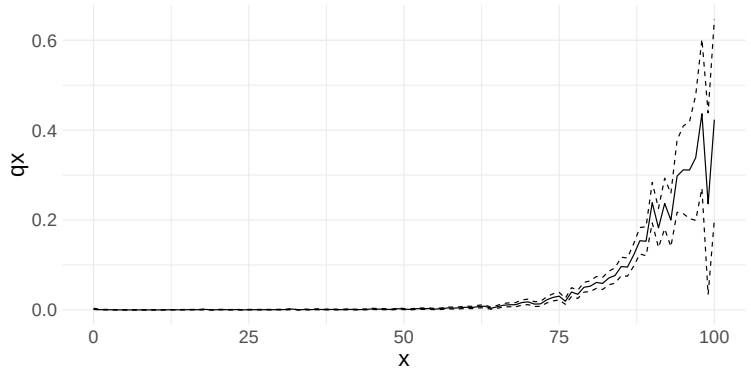

Figure 1: Probabilités de décès en fonction de l'âge

avec $q_{1-\alpha/2}$ le quantile de la loi normale d'ordre $1-\alpha/2$.

(c) (1 point) Nous décidons d'utiliser le modèle suivant

$$\text{logit}(q_x) = \log\left(\frac{q_x}{1-q_x}\right) = \alpha + \beta x, \ x \in \{0, \ldots, 100\}. \tag{1}$$

De quelle type de procédure s'agit-il? Pourquoi utiliser une telle procédure? Votre réponse peut s'appuyer sur la Figure 1.

Solution:

- Il s'agit d'une procédure de lissage paramétrique (0.5 point)
- L'idée est compenser la volatilité des estimations des probabilités de décès aux grands âge due à une faible exposition. (0.5 point)

(d) (1 point) La Figure 2 montre le logit des probabilités de décès en fonction de l'âge.

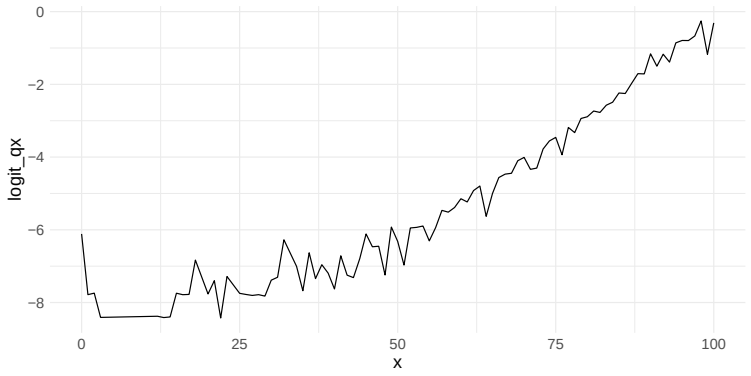

Figure 2: logit des probabilités de décès en fonction de l'âge

Pensez vous que le modèle (1) est adapté? Quelle information risque-t-on de rater?

Solution:

- On note la tendance linéaire du logit des probabilités de décès en fonction de l'âge en particulier à partir de $40 - 50$ ans. (0.5 point)
- On ne pourra pas capter la mortalité infantile qui risque en plus de poluer l'estimation de α et β. (0.5 point)

(e) (2 points) Après application de la méthode précédente, nous obtenons les probablités de décès en pointillé sur la Figure 3. Que pensez-vous de cet ajustement? Comment quantifier

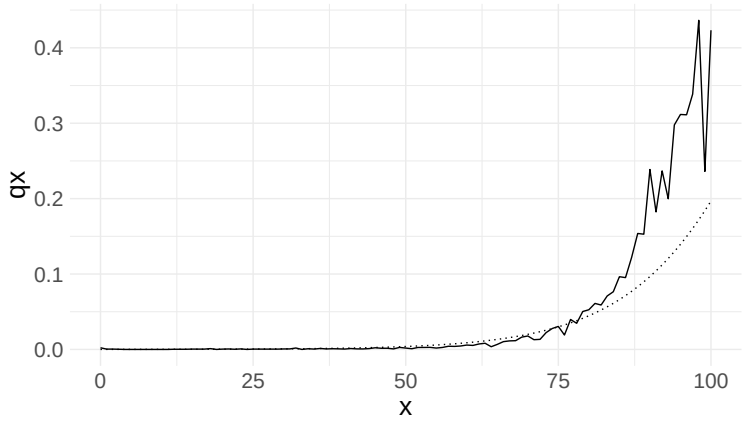

Figure 3: Probabilités de décès en fonction de l'âge après ajustement via le modèle logistique

la qualité de l'ajustement des probabilités de décès? Comment améliorer cet ajustement? (On peut s'appuyer sur la Figure 2 pour justifier sa réponse).

Solution:

1. L'ajustement est correct même si on constate une sous estimation des probabilités de décès aux grands âges. (0.5 point)
2. Il faut mesurer la fidélité des taux lissés par rapport aux taux bruts, par exemple vérifier que les taux lissés tombent dans l'intervalle de confiance des taux bruts. (0.5 points) On peut effectuer un test des signes pour vérifier qu'on ne surestime ou ne sous estime pas systématiquement les taux bruts. (0.5 point)
3. Il faudrait probablement restreindre l'interval des âges pour l'ajustement du modèle. Par exemple retirer les âges inférieurs à 40 ans et supérieur à 80 ans. Ce modèle logistique peut être utiliser pour extrapoler les probabilités de décès (méthode de Kannisto vu en cours). (0.5 point)

3. Nous disposons de données sur la durée des arrêts de travail dans une entreprise. Le tableau 3 donne un aperçu des données.

time	status	age	gender
5.01	1	25	1
1.14	0	50	0
7.83	1	60	0
6.75	1	25	1
5.86	0	53	0
0.27	1	55	1

Table 1: Durée des arrêts de travail dans l'entreprise

Les variables sont décrites ci-dessous:

- `time`: Durée de l'arrêt de travail
- `status`: 0= "en cours" et 1 = "terminé". Un 0 équivaut à une donnée censurée à droite.
- `age`: âge du salarié ou de la salariée
- `gender`: genre du salarié ou de la salariée

(a) (2 points) Le jeu de données est stocké sous la forme d'un *data frame* nommé `df_incap`. Nous débutons l'analyse avec le code suivant

```
fit <- survfit(Surv(time, status) ~  gender, data = df_incap)
ggsurvplot(fit,
           pval = TRUE, conf.int = TRUE,
           ggtheme = theme_bw(),
           linetype = "strata"
)
```

Le résultat est donné sur la Figure 4. A quoi correspondent les différents éléments visibles sur le graphique? Vous devez préciser les méthodologies statistiques employées (les formules ne sont pas nécessaires) et donner votre interprétation de ce graphique.

Solution:

- Nous avons les fonctions de survies de la durée des arrêts de travail au sein des populations de salariés discriminées par la variable `gender`. En présence de données censurés à droite, l'estimation est obtenue via l'estimateur de Kaplan-Meieir. La valeur estimée est entouré d'une bande qui correspond aux intervalles de confiances. (0.5 point) Nous constatons que les deux courbes de survies tombe dans l'intervalle de confiance de l'autre courbe, cela donne une indiquation quant à leur similarité statistique. (0.5 point)
- L'élément $p = 0.55$ est la *p-value* du test de log rang testant l'hypothèse de similarité des courbes de survie (0.5 point). Cette valeur "élevée" (supérieur à un seuil de significativité, 0.05 par exemple) indique qu'il n'est pas possible de rejeter H_0 et donc qu'il n'y a pas de différence statistiquement significative entre les deux courbes (0.5 point).

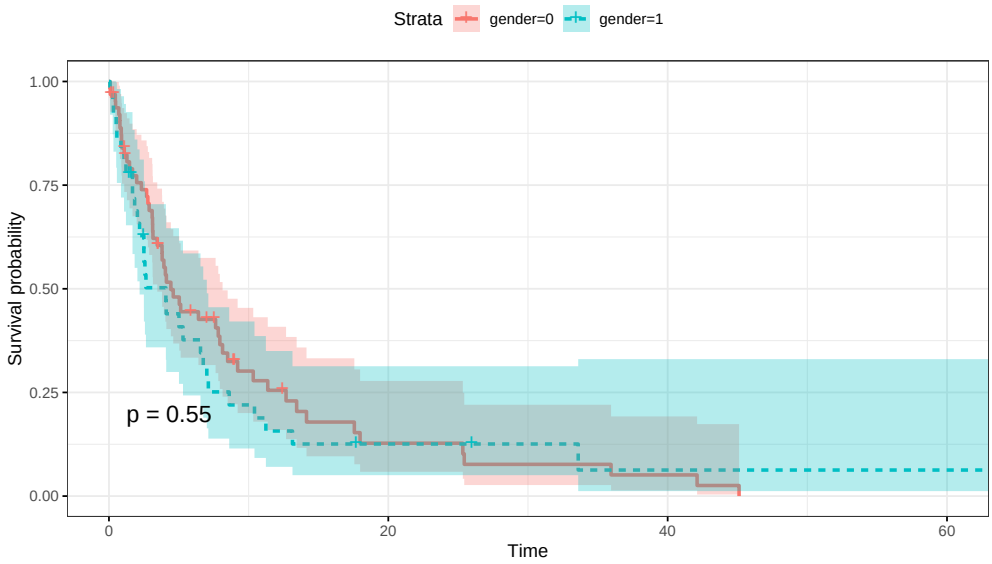

Figure 4: ?

(b) (3 points) Nous poursuivons avec le code suivant

```
res.cox <- coxph(Surv(time, status) ~ age + gender, data = df_incap)
res.cox
```

Nous obtenons dans la console le résultat de la Figure 5. Quel est le modèle considéré?

```
Call:
coxph(formula = Surv(time, status) ~ age + gender, data = df_incap)

            coef exp(coef) se(coef)     z        p
age     0.034812  1.035425 0.009882 3.523 0.000427
gender1 0.139110  1.149251 0.230597 0.603 0.546334

Likelihood ratio test=12.22  on 2 df, p=0.00222
n= 100, number of events= 83
```

Figure 5: ?

Rappeler les hypothèses avec une formule adaptée à l'étude de la durée des arrêts de travail considérées dans cet exercice. Quelles conclusions retirez-vous de ces résultats?

Solution:

- Le modèle de risque proportionel de Cox est utilisé pour analyser l'impact des covariables `age` et `gender` sur la durée des arrêts de travail (0.5 point)
- Le modèle de Cox définit la fonction de hasard par

$$h(t) = h_0(t) \exp(\beta_1 \cdot age + \beta_2 \cdot \mathbb{I}_{gender=1})$$

(1 points)

- La durée des arrêts de travail augmente avec l'âge car exp(coef) > 1, de plus le test de significativité du coefficient renvoie une p-valeur très inférieure à 0.05 (0.5 point)
- La durée des arrêts de travail augmente si gender = 1 car exp(coef) > 1, en revanche le test de significativité du coefficient renvoie une p-valeur très supérieure à 0.05 ce qqui indique que l'impact de la variable gender n'est pas suffisante pour être significative. Cela corrobore les résultats obtenus dans la question précédente (0.5 point)
- Dans son ensemble le modèle est pertinent comme l'indique le test du rapport de vraisemblance dont la pvaleur est très inférieur à 0.05 (0.5 point)