
Travaux dirigés : modèles de durée
Séance n°3

Exercice 1 Modèle binomiale et ajustement logistique.

On considère un modèle pour la durée de la vie humaine en temps discret pour les âges entiers $x \in \{x_{\min}, \dots, x_{\max}\}$. On suppose que le nombre de décès D_x à l'âge x suit une loi Binomiale $\mathcal{B}(n_x, q_x)$, où n_x correspond à l'effectif observé à l'âge x .

1. Pour une population donnée où n_x et d_x peuvent être calculés, rappeler l'estimateur du maximum de vraisemblance de q_x . Ces estimateurs des probabilités conditionnelles \hat{q}_x sont généralement appelés probabilités de décès brutes.
2. On se place à présent dans le cadre d'un modèle linéaire généralisé, et on suppose que q_x s'écrit tel que

$$q_x = \frac{\exp(\eta_x)}{1 + \exp(\eta_x)},$$

avec le prédicteur linéaire $\eta_x = \sum_{s=0}^p \beta_s x^s$, où $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ est un vecteur de paramètres à estimer.

Écrire la vraisemblance du modèle et indiquer comment obtenir un estimateur du vraisemblance pour $\boldsymbol{\beta}$.

3. Comment contrôler la qualité de l'ajustement réalisé.
4. Une autre spécification classique consiste à analyser le modèle de régression suivant pour chaque â

$$y_x = \sum_{s=0}^p \alpha_s x^s + \epsilon_x,$$

avec ϵ_x un bruit blanc et $y_x = \text{logit}(\hat{q}_x)$. Comparer le biais des estimateurs obtenus avec cette approche. Commenter ce résultat.

Exercice 2 Analyse du modèle à hasard proportionnel de Cox.

On considère des données de survie (avec censure aléatoire à droite indépendante et non-informative). Elles sont influencées par un certain nombres de covariables observées \mathbf{X} (non touchées par la censure). Ces données sont analysées au moyen d'un modèle à hasard proportionnel, pour lequel une fonction de base non spécifiée est utilisée (modèle semi-paramétrique de Cox).

1. En notant $\boldsymbol{\beta}$ le vecteur de paramètres du modèle, rappeler l'expression de la fonction de hasard $h(t | \mathbf{X}; \boldsymbol{\beta})$ d'un modèle à hasard proportionnel où $h_0(t)$ est la fonction hasard de base. En déduire l'expression de la fonction de survie, puis celle des probabilités de décès conditionnelles $q(t | \mathbf{X}; \boldsymbol{\beta})$.

2. On considère deux individus i et j de covariables \mathbf{X}_i et \mathbf{X}_j . Commenter le ratio de leur fonction de hasard.
3. On analyse au moyen d'un tel modèle la durée de vie résiduelle d'individus malades en fonction du sexe* et de l'âge auquel la maladie s'est développée. Les résultats obtenus sous le logiciel R sont les suivants.

```

Call:
coxph(formula = Surv ~ age + sex, data = t)

n= 228, number of events= 165

      coef exp(coef)   se(coef)      z Pr(>|z|)
age  0.017045  1.017191  0.009223  1.848  0.06459 .
sex -0.513219  0.598566  0.167458 -3.065  0.00218 **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

exp(coef) exp(-coef) lower .95 upper .95
age     1.0172      0.9831    0.9990    1.0357
sex     0.5986      1.6707    0.4311    0.8311

Concordance= 0.603  (se = 0.026 )
Rsquare= 0.06  (max possible= 0.999 )
Likelihood ratio test= 14.12  on 2 df,  p=0.0008574
Wald test            = 13.47  on 2 df,  p=0.001187
Score (logrank) test = 13.72  on 2 df,  p=0.001048

```

Commenter ces résultats (significative et interprétation des effets).

4. Rappeler comment peut-être analysée l'hypothèse de proportionnalité. Commenter les résultats ci-dessous et la Figure 1

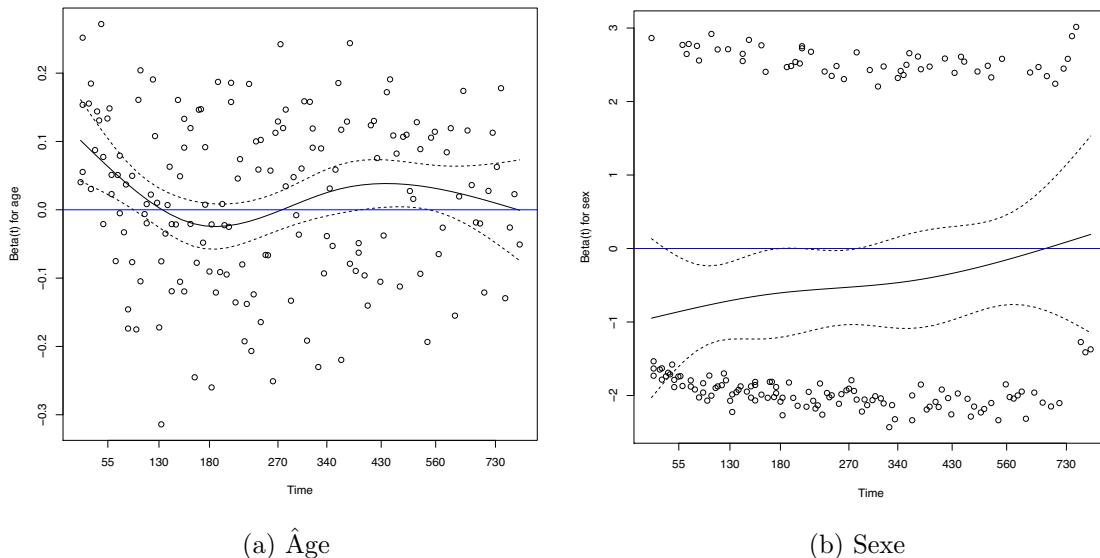


FIGURE 1 – Analyse graphique des résidus de

*. Codage : 1 pour les hommes et 2 pour les femmes.

	rho	chisq	p
age	-0.0275	0.129	0.719
sex	0.1236	2.452	0.117
GLOBAL	NA	2.651	0.266

Exercice 3 Modèle de Hannerz.

De nombreux modèles paramétriques ont été proposés dans la littérature. Beaucoup sont souvent peu paramétrés et ne sont applicables qu'à certaines tranches d'âge de la population. Le modèle de Hannerz * tente de rendre compte de la mortalité féminine suédoise dans son ensemble. Ce modèle s'écrit en fonction de 5 paramètres (a_0, \dots, a_3, c) strictement positifs

$$G(x) = \ln\left(\frac{1 - S(x)}{S(x)}\right) = a_0 - \frac{a_1}{x} + \frac{a_2}{2}x^2 + \frac{a_3}{c}e^{cx}.$$

1. Montrer que $g(x) = \frac{dG(x)}{dx} = \frac{h(x)}{1 - S(x)} = a_1x^{-2} + a_2x + a_3e^{cx}$.
2. Le terme a_1x^{-2} permet de capturer le pic de mortalité aux âges jeunes. Expliquer pourquoi ?
3. Sur la population féminine, le terme a_2x correspond à un accroissement de $g(x)$ sur la période comprise entre 16 et 64 (période d'activité) dans le modèle original de Hannerz. Cette spécification vous semblerait-elle adaptée à la population masculine ?
4. Que pensez-vous du dernier terme ?
5. Bien souvent les données disponibles en assurance ne permettent pas de construire un table complète. Un actuaire cherchera donc plutôt à utiliser une logique de positionnement. Le modèle relationnel construit sur le modèle de Hannerz prend la forme

$$G(x) - G^{\text{ref}}(x) = b_0 - \frac{b_1}{x} + \frac{b_2}{2}x^2 + \frac{b_3}{c}e^{cx}.$$

Cette relation vous paraît-elle facilement interprétable ?

6. Comment estimeriez vous ce modèle ?

Exercice 4 Modélisation prospective de la mortalité.

Le risque de longévité d'un régime de retraite est usuellement analysé au moyen d'un modèle de mortalité prospective. Ces analyses s'appuient généralement sur les données d'une population nationale. On suppose que la fonction de hasard à l'âge x et pour la date t , notée $\mu_x(t)$ est constante sur un carré $[x, x+1] \times [t, t+1]$ du diagramme de Lexis.

On s'intéresse au modèle de Lee-Carter

$$\ln(\hat{\mu}_x(t)) = \alpha_x + \beta_x \kappa_t + \epsilon_{xt},$$

avec ϵ_{xt} des erreurs centrées, indépendantes et de variance σ^2 .

1. Interpréter les coefficients du modèle.

*. Hannerz, H. (2001). Presentation and derivation of a five-parameter survival function intended to model mortality in modern female populations. Scandinavian Actuarial Journal, 2001(2), 176-187.

2. Les contraintes

$$\sum_{x=x_{\min}}^{x_{\max}} \beta_x = 1 \text{ et } \sum_{x=t_{\min}}^{t_{\max}} \kappa_t = 0,$$

sont appliquées sur les coefficients. Expliquer pourquoi ?

3. Poser le problème d'optimisation considéré dans le modèle traditionnel de Lee-Carter pour déterminer les paramètres (α, β, κ) . Quel est l'inconvénient majeur de cette approche ?
4. Proposer une reformulation du modèle pour corriger cet inconvénient.
5. Dans le cadre de cette seconde approche, nous avons tracé les résidus de la déviance (cf. Figure 2) du modèle en fonction de l'âge et de l'année calendaire à partir des données nationales correspondant à la mortalité des hommes en Angleterre et au Pays de Galles. Quelle limite importante voyez-vous apparaître ? Comment la corrigeeriez-vous ?

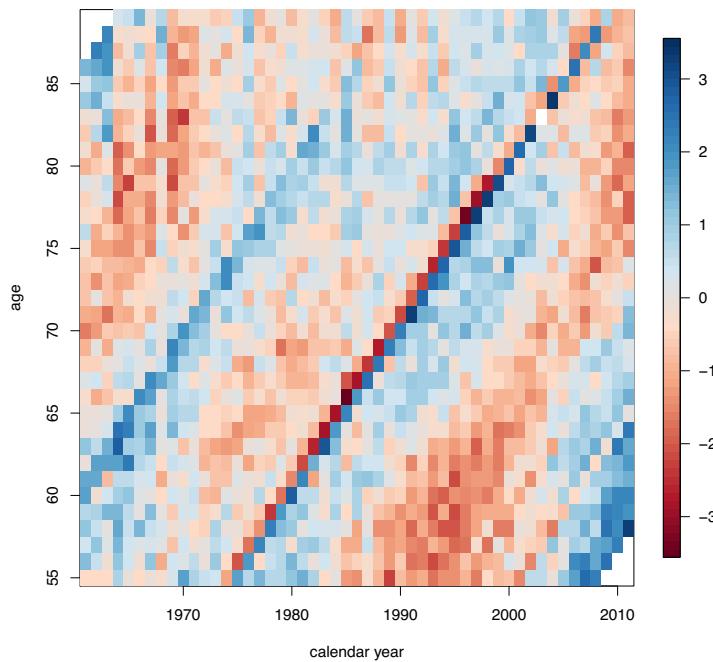


FIGURE 2 – Analyse graphique des résidus de la déviance en fonction de l'âge et de l'année calendaire.

6. Expliquer comment procéder à la projection des taux de mortalité pour des dates t futures ?

Exercice 5 Modèle à risques concurrents.

Soit T la durée de vie d'un individu que l'on suppose soumise à K causes de sortie (ex : causes de décès). On définit la *fonction de hasard spécifique au risque i*

$$h^{(i)}(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t, V = i | T > t),$$

où V est la variable aléatoire qui désigne la cause de sortie.

1. Donner l'expression de :

- la fonction de hasard h (associée à T) en fonction des $h^{(i)}$;
- la fonction de survie S (associée à T) en fonction des $h^{(i)}$;
- la *fonction d'incidence cumulée* $F^{(i)}(t) = \mathbb{P}(T \leq t, V = i)$;
- le taux de décès entre les dates t et $t + 1$ pour la cause $V = i$.

2. On introduit les durées de vie latentes par cause T_1, T_2, \dots, T_K . Exprimer $S(t)$ en fonction des fonctions de survie de chacune des lois latentes en faisant l'hypothèse qu'elles sont indépendantes. Que peut-on en déduire s'agissant de l'expression des taux de hasard de chacune de ces lois latentes ?

3. En notant $S_{(T_1, \dots, T_K)}(t_1, \dots, t_K)$ la fonction de survie jointe de T_1, T_2, \dots, T_K et sans supposer que ces lois sont nécessairement indépendantes, montrer que

$$h^{(i)}(t) = -\frac{1}{S(t)} \left. \frac{\partial S_{(T_1, \dots, T_K)}(t_1, \dots, t_K)}{\partial t_i} \right|_{t_1, \dots, t_K=t}.$$

4. Pour $K = 2$, nous faisons à présent l'hypothèse que les durées de vie latentes sont corrélées et reliées par la fonction de survie jointe

$$S_{T_1, T_2}(t_1, t_2) = (1 + \theta(\lambda_1 t_1 + \lambda_2 t_2))^{-\frac{1}{\theta}},$$

avec $\lambda_1, \lambda_2, \theta \geq 0$. Fournir l'expression de la fonction de hasard $h^{(i)}(t)$ pour $i = 1, 2$, puis celle de $F^{(i)}(t)$.

Exercice 1.

© Théo Jalabert

$$1) \hat{q}_x = \frac{D_x}{n_x}$$

1

Exercise 1 Modèle binomiale et ajustement logistique.

On considère un modèle pour la durée de la vie humaine en temps discret pour les âges entiers $x \in \{x_{\min}, \dots, x_{\max}\}$. On suppose que le nombre de décès D_x à l'âge x suit une loi Binomiale $\mathcal{B}(n_x, q_x)$, où n_x correspond à l'effectif observé à l'âge x .

- Pour une population donnée où n_x et d_x peuvent être calculés, rappeler l'estimateur du maximum de vraisemblance de q_x . Ces estimateurs des probabilités conditionnelles \hat{q}_x sont généralement appelés probabilités de décès brutes.
- On se place à présent dans le cadre d'un modèle linéaire généralisé, et on suppose que q_x s'écrit tel que

$$q_x = \frac{\exp(\eta_x)}{1 + \exp(\eta_x)},$$

avec le prédicteur linéaire $\eta_x = \sum_{s=0}^p \beta_s x^s$, où $\beta = (\beta_0, \dots, \beta_p)$ est un vecteur de paramètres à estimer.

Écrire la vraisemblance du modèle et indiquer comment obtenir un estimateur du vraisemblance pour β .

- Comment contrôler la qualité de l'ajustement réalisé.
- Une autre spécification classique consiste à analyser le modèle de régression suivant pour chaque \hat{q}_x

$$y_x = \sum_{s=0}^p \alpha_s x^s + \epsilon_x,$$

avec ϵ_x un bruit blanc et $y_x = \logit(\hat{q}_x)$. Comparer le biais des estimateurs obtenus avec cette approche. Commenter ce résultat.

$$\hat{q}_x = \frac{e^{\eta_x}}{1+e^{\eta_x}}$$

Quest°: $(f(\beta) = \sum_{x=x_{\min}}^{x_{\max}} \frac{E_x}{\hat{q}_x(1-\hat{q}_x)} (\hat{q}_x - q_x(\beta))^2 \rightarrow \text{Déviance}$

→ audio

- L'estimateur "naturel" des taux bruts est

$$\hat{q}_x = \frac{d_x}{n_x}$$

On retrouve aisément ce résultat en considérant l'estimateur du maximum de vraisemblance du paramètre de fréquence d'une loi binomiale.

- En utilisant des notations évidentes, la vraisemblance du modèle s'écrit en regroupant par âge entier

$$\mathcal{L}(\beta) = \prod_{x=x_{\min}}^{x_{\max}} \binom{n_x}{d_x} q_x^{d_x} (1-q_x)^{n_x-d_x}$$

1

Ainsi, la log-vraisemblance du modèle est égale (à une constante près) à

$$\sum_{x=x_{\min}}^{x_{\max}} d_x \eta_x - n_x \ln(1 + \exp(\eta_x)).$$

En dérivant par rapport à β_s , le système à résoudre pour obtenir les paramètres du modèle est, pour tout $s = 0, \dots, p$

$$\sum_{x=x_{\min}}^{x_{\max}} x^s \left(d_x - n_x \frac{\exp(\eta_x)}{1 + \exp(\eta_x)} \right) = \sum_{x=x_{\min}}^{x_{\max}} x^s (d_x - q_x n_x) = 0,$$

La première relation pour $s = 0$ permet d'assurer la reproduction du nombre total de décès par le modèle, i.e. $\sum_{x=x_{\min}}^{x_{\max}} d_x = \sum_{x=x_{\min}}^{x_{\max}} \hat{q}_x n_x$. En posant le résidu

$$r_x = d_x - n_x \frac{\exp(\eta_x)}{1 + \exp(\eta_x)},$$

les équations suivantes apparaissent comme une relation d'orthogonalité entre les résidus et les variables x^s . Ces équations peuvent être résolues numériquement en appliquant un algorithme de Newton-Raphson.

- Comme pour un modèle linéaire généralisé classique, la qualité du modèle peut être analysée par le biais de :

— la déviance (avec $\ln(\mathcal{L}(d))$), la log-vraisemblance du modèle saturé, i.e. pour lequel les observations sont exactement prédites, et $\ln(\mathcal{L}(\hat{d}))$), la log-vraisemblance du modèle ajusté)

$$\begin{aligned} \text{Dev}(d, \hat{d}) &= \sum_{x=x_{\min}}^{x_{\max}} \text{Dev}_x(d, \hat{d}) = 2 \left(\ln(\mathcal{L}(d)) - \ln(\mathcal{L}(\hat{d})) \right) \\ &= 2 \sum_{x=x_{\min}}^{x_{\max}} d_x \ln \left(\frac{d_x}{n_x \hat{q}_x} \right) - (n_x - d_x) \ln \left(\frac{n_x - d_x}{n_x - n_x \hat{q}_x} \right) \end{aligned}$$

— un test d'adéquation du chi-2 de Pearson

$$\chi^2 = \sum_{x=x_{\min}}^{x_{\max}} \frac{(d_x - n_x \hat{q}_x)^2}{n_x \hat{q}_x (1 - \hat{q}_x)}$$

— le pseudo- R^2

$$\text{pseudo-}R^2 = 1 - \frac{\sum_{x=x_{\min}}^{x_{\max}} (d_x/n_x - \hat{q}_x)^2}{\sum_{x=x_{\min}}^{x_{\max}} d_x/n_x - (n^{-1} \sum_{x=x_{\min}}^{x_{\max}} d_x/n_x)^2}$$

Exercice 4:

© Théo Jalabert



Lee Carter : $\ln(\mu_l(x,t)) = \alpha_x + \beta_x k_t + \varepsilon_{xt}$

- 1) α_x : niveau moyen du log des taux constraintés à chaque âge.
Généralement croissants