

Modèles de durée / Examen du 13 février 2006

Durée 1,5h – aucun document n'est autorisé

Corrigé

Problème (modèle logistique)

Dans le cadre de la construction d'une table d'expérience pour des contrats en cas de décès, on s'intéresse ici à l'ajustement de taux de mortalité bruts \hat{q}_x par un modèle logistique. La fonction logistique est par définition $\mathbf{lg}(x) = \ln\left(\frac{x}{1-x}\right)$.

Question n°1

Donner l'intervalle de définition de $x \rightarrow \mathbf{lg}(x)$ et calculer les dérivées première et seconde, ainsi que l'inverse. Que peut-on en conclure ?

La fonction logistique est définie sur $]0,1[$; elle est croissante sur cet intervalle :

$$\frac{d}{dx} \mathbf{lg}(x) = \frac{1}{x(1-x)}.$$

On a par ailleurs :

$$\frac{d^2}{dx^2} \mathbf{lg}(x) = \frac{2x-1}{x^2(1-x)^2}.$$

Sur l'intervalle $\left]0, \frac{1}{2}\right[$ la fonction $\mathbf{lg}(x)$ est donc concave (et convexe sur $\left[\frac{1}{2}, 1\right[$).

Question n°2

On fait l'hypothèse que l'on a estimé le taux de décès à l'âge x par \hat{q}_x supposé dans biais ; on se propose d'effectuer un ajustement des $\hat{y}_x = \mathbf{lg}(\hat{q}_x)$. Prouvez que l'estimateur \hat{y}_x est biaisé pour estimer $\mathbf{lg}(q_x)$ et indiquez le sens du biais (on pourra supposer que les taux à estimer sont inférieurs à $\frac{1}{2}$). Qu'en concluez-vous ?

On rappelle l'inégalité de Jensen pour une fonction convexe : $f(EX) \leq Ef(X)$.

D'après l'inégalité de Jensen, dans une zone où les taux de décès sont petits, et si on a estimé le taux de décès par \hat{q}_x supposé dans biais, alors :

$$E \lg(\hat{q}_x) \leq \lg(E(\hat{q}_x)) = \lg(q_x)$$

En d'autres termes, les logits empiriques ainsi obtenus sont biaisés négativement (ils sous-estiment les vrais logits) ; mais comme la fonction $\lg(x)$ (et son inverse) est croissante, en sous-estimant les logits théoriques, cette démarche sous-estime les taux de décès théoriques.

La conclusion est inverse pour des taux de sortie supérieurs à $\frac{1}{2}$.

Dans le cadre d'un ajustement des $\hat{y}_x = \lg(\hat{q}_x)$, on obtient les taux de décès par la transformation inverse $y \rightarrow \frac{e^y}{1+e^y}$; la présence d'exponentielles dans cette expression conduit à une amplification importante du biais d'estimation évoqué ci-dessus. Ainsi, dans le cas d'un risque décès, un modèle d'ajustement des logits des taux de décès conduit à sous-estimer dans des proportions qui peuvent être importantes (typiquement de 5 à 10%) les taux de décès.

Les modèles utilisant les logits des taux de décès doivent donc être utilisés avec prudence dans le cas d'un risque en cas de décès. Ils intègrent au contraire une marge de sécurité dans le cas d'un risque en cas de vie.

Question n°3

On suppose une relation affine entre l'âge et le logit du taux de décès ; en d'autres termes on suppose que $\lg(q_x) = a + bx$ pour des paramètres a et b inconnus. Proposez un modèle simple et « naturel » pour estimer les paramètres a et b et décrivez en les principales propriétés.

Le modèle de base d'ajustement logistique part du constat que sur une large plage le logit des taux de décès présente une tendance linéaire ; on propose alors la modélisation suivante :

$$\lg(\hat{q}_x) = a + bx + \varepsilon$$

où ε est un bruit gaussien iid ; on régresse donc simplement les logits des taux de décès sur l'âge.

Question n°4

Montrez que la paramétrisation $\lg(q_x) = a + bx$ peut s'écrire $q_x = \frac{ce^{dx}}{1+ce^{dx}}$ avec des paramètres c et d que l'on précisera en fonction de a et b .

Déterminez la fonction de survie et la fonction de hasard de ce modèle.

La transformation inverse du logit étant $y \rightarrow \frac{e^y}{1+e^y}$, le modèle $\text{lg}(q_x) = a + bx$ s'écrit de manière équivalente :

$$q_x = \frac{ce^{dx}}{1+ce^{dx}}$$

en posant $c = e^a$ et $d = b$. Une approche alternative à la régression linéaire $\text{lg}(\hat{q}_x) = a + bx + \varepsilon$ consiste donc à effectuer une estimation par maximum de vraisemblance

dans le modèle paramétrique $q_x = \frac{ce^{dx}}{1+ce^{dx}}$. Cette approche évite *a priori* l'effet de sous estimation des taux de mortalité associée à l'approche par régression linéaire, le taux de décès étant la variable modélisée (mais l'estimateur du maximum de vraisemblance n'a toutefois pas de raison d'être sans biais).

La détermination de la fonction de survie et de la fonction de hasard, liées l'une à l'autre part

la relation $S(t) = \exp\left(-\int_0^t \mu(s) ds\right)$ nécessite de faire des hypothèses. En effet, la relation

$q(x) = 1 - \frac{S(x+1)}{S(x)}$ conduit dans le cas général à la contrainte sur la fonction de hasard :

$$-\ln(1-q_x) = \int_x^{x+1} \mu(s) ds$$

Dans le modèle discret spécifié jusqu'alors x est *a priori* entier. Il faut donc une règle de passage du temps discret au temps continu. On peut utiliser différentes approches (Balducci, constance des taux de hasard par morceau, etc.). Si on choisit l'hypothèse de constance de la fonction de hasard entre deux valeurs entières, on trouve que la fonction de hasard est une fonction en escalier avec aux points entiers :

$$\mu_x = \frac{cde^{dx}}{1+ce^{dx}}$$

Question n°5

En supposant que l'estimateur \hat{q}_x est déterminé dans le cadre d'un modèle binomial et que les effectifs sont suffisants pour utiliser l'approximation normale, écrivez la log-vraisemblance du modèle.

On rappelle que la densité d'une variable gaussienne d'espérance m et de variance σ^2 s'écrit : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$.

En partant de la remarque que la loi du taux brut est de la forme $N(\hat{q}_x; \sigma^2(\theta)) = \frac{q_x(\theta)(1-q_x(\theta))}{N_x}$, on en déduit $L(\theta) = \prod_x \frac{1}{\sigma(\theta)\sqrt{2\pi}} \cdot \exp\left(\frac{1}{2} \frac{(q_x(\theta)-\hat{q}_x)^2}{\sigma^2(\theta)}\right)$, d'où :

$$\ln(L(\theta)) = \sum_x \ln\left(\frac{1}{\sigma(\theta)\sqrt{2\pi}}\right) - \sum_x \frac{1}{2} \frac{(q_x(\theta)-\hat{q}_x)^2}{\sigma^2(\theta)}$$

Dans le cas présent, $\theta = (c, d)$ et $q_x(\theta) = \frac{ce^{dx}}{1+ce^{dx}}$.

Question n°6

En remplaçant la variance théorique par la variance estimée dans la formule précédente, proposez une approximation de la log-vraisemblance déterminée à la question précédente qui ramène la recherche du maximum de vraisemblance à un problème de moindres carrés non linéaires.

Lorsqu'on remplace la variance théorique par la variance estimée, la maximisation de la vraisemblance est alors équivalente à la minimisation de :

$$\sum_x \frac{1}{2} \frac{(q_x(\theta)-\hat{q}_x)^2}{\hat{\sigma}^2} = \sum_x \frac{N_x}{\hat{q}_x(1-\hat{q}_x)} (q_x(\theta)-\hat{q}_x)^2$$

Le problème est ainsi ramené à un problème de moindres carrés pondérés dans le cas non linéaire ; il peut être résolu numériquement dans la plupart des logiciels statistiques spécialisés.

Question n°7

Proposez une méthode de détermination des valeurs initiales des paramètres c et d que l'on pourrait utiliser dans un algorithme numérique.

On peut simplement déterminer a et b à l'aide de la régression linéaire, puis obtenir ensuite des valeurs initiales des paramètres c et d à partir des formules de passage à partir de a et b . Les estimateurs ainsi obtenus pour c et d sont biaisés, mais *a priori* suffisants pour initialiser un algorithme numérique.