REVUE DE STATISTIQUE APPLIQUÉE

J. O'QUIGLEY

Intervalles de confiance pour les estimations des courbes de survie à partir du modèle de Cox

Revue de statistique appliquée, tome 32, n° 1 (1984), p. 39-45

<http://www.numdam.org/item?id=RSA_1984__32_1_39_0>

© Société française de statistique, 1984, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

INTERVALLES DE CONFIANCE POUR LES ESTIMATIONS DES COURBES DE SURVIE A PARTIR DU MODELE DE COX

J. O'QUIGLEY

Unité de Recherches Statistiques de l'INSERM,
16, avenue Paul Vaillant Couturier, 94807 Villejuif Cedex France

1. INTRODUCTION

Le modèle de COX [1] qui permet d'étudier les données de survie (éventuellement censurées) ne suppose pas la connaissance d'une fonction paramétrique des risques instantanés de décès. Du fait de ce caractère non paramètrique et de la robustesse qui en résulte cette méthode est très utilisée en pratique pour tester la signification de facteurs de risque. Mais on est parfois en droit d'attendre mieux d'un modèle que la formulation d'hypothèses et les tests qui en découlent. A titre d'exemple, dans certains cas il pourrait être utile de construire un indice pronostique à partir de différents facteurs quantitatifs ou qualitatifs connus pour influencer la survie. Le but d'un modèle serait alors non pas de déceler des traitements faiblement actifs, compte tenu d'autres covariables, mais de quantifier les effets relatifs de chaque variable, ce qui permet d'assigner à chaque sujet un score correspondant en quelque sorte à sa probabilité de survie.

Dans le langage statistique ce problème est considéré comme un problème d'estimation. Il faut d'abord vérifier le modèle et la condition préalable, avant qu'on puisse utiliser les indices ainsi trouvés, est que le modèle s'adapte bien aux données, ce qui peut être vérifié par différentes méthodes [2, 3, 4, 5].

Enfin si le modèle se révèle approprié il faudra prendre en compte les fluctuations d'échantillonnage. TSIATIS [6] a calculé les courbes de survie à partir du modèle de COX, ses résultats étant les mêmes que ceux de BRESLOW [7] dans le cas où il n'y a pas d'exaequos dans l'échantillon pour les durées de survie. Le calcul est effectué en remplaçant les paramètres par leur estimateurs.

La prise en compte des variances de ces estimateurs a amené TSIATIS [6] à calculer un intervalle de confiance pour les risques instantanés de décès cumulés dans le cas où il n'y a pas d'exaequos et où les covariables (éventuellement centrées) prennent la valeur zéro.

Le but de ce travail est de généraliser ce dernier résultat. Cette généralisation consiste :

1) à calculer les intervalles de confiance pour les fonctions de survie estimées,
2) à permettre aux covariables de prendre des valeurs quelconques,
3) à prendre en compte la présence éventuelle d'exaequos parmi les durées de survie.

A titre d'illustration, un exemple, concernant la survie des malades atteints d'un cancer de l'estomac, est montré où les intervalles de confiance à 95 % sont calculés pour les courbes de survie de trois groupes.

Mots-Clés : **Modèle de Cox, Intervalles de confiance.**

2. LE MODELE DE COX ET LES RESULTATS DE TSIATIS

Soit T une variable aléatoire qui représente la durée de survie. Trois fonctions statistiques sont utiles pour étudier la distribution de T.

$$S(t) = \text{pr}(T > t) \tag{1}$$

$$f(t) = \lim_{\Delta t \to 0} \frac{\text{pr}(t < T < t + \Delta t)}{\Delta t} \tag{2}$$

$$\lambda(t) = \lim_{\Delta t \to 0} \text{pr}(t < T < t + \Delta t \mid T > t)/\Delta t \tag{3}$$

Elles sont d'ailleurs liées les unes aux autres par :

$$\lambda(t) = -\frac{d}{dt}(\log S(t)) = f(t)/S(t) \tag{4}$$

$$S(t) = \exp\left[-\int_0^t \lambda(u)\,du\right] \tag{5}$$

$$f(t) = -\frac{dS(t)}{dt} = \lambda(t)\exp\left[-\int_0^t \lambda(u)\,du\right] \tag{6}$$

Cox [1] a proposé comme modèle de survie :

$$\lambda(t, x_i) = \lambda_0(t)\exp(\beta' x_i) \tag{7}$$

où t est la réalisation d'une variable aléatoire, T, la durée de survie et x_i un vecteur de p covariables pour le sujet i. Donc

$$x_i' = (x_{i1}, x_{i2}, \ldots . x_{ip})$$

$$\beta' = (\beta_1, \beta_2, \ldots . \beta_p)$$

Les paramètres β sont fixes mais, dans presque tous les cas, inconnus. La formule (7) exige que chaque covariable agisse d'une façon multiplicative sur $\lambda(t, x_i)$ la fonction de risque instantanée. C'est ainsi qu'on peut utiliser la vraisemblance partielle [8] et par la suite en induire les estimations de β sans qu'on sache la forme de $\lambda_0(t)$. En effet pour un échantillon de n individus où $t_1 < t_2 < \ldots < t_{k^*}$ les temps de mort observés ($t_{k^*+1} = \infty$), la vraisemblance s'écrit comme suit [9] :

$$\prod_{i=1}^{k^*} p(A_i \mid B_i) \left\{ \prod_{i=2}^{k^*} p(B_i \mid B_{i-1}, A_{i-1}) \times p(B_1) \times p(B_{k^*+1}) \right\}$$

où A_i est l'évènement : l'individu i meurt à t_i et B_i est l'évènement : une mort s'est produite à t_i et telle configuration de morts et de données censurées s'est produite avant t_i. Le premier terme de cette expression ne fait pas intervenir $\lambda_0(t)$ et se simplifie en

$$V(\beta) = \prod_{i=1}^{k^*} \left\{ \frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \right\}$$

où $R(t_i)$ désigne l'ensemble des sujets "à risque" à t_i, à savoir des sujets ni censurés ni morts juste avant t_i. Lorsqu'il existe une valeur $\hat{\beta}$ de sorte que $V(\hat{\beta}) > V(\beta^*)\ \forall \beta^*$ on appelle $\hat{\beta}$ l'estimateur du maximum de vraisemblance partielle. On remarque aussitôt une propriété fort attrayante de $\hat{\beta}$, c'est-à-dire que $\hat{\beta}$ dépend bien sûr, de l'échantillon mais non pas de la forme de $\lambda_0(t)$. Inutile donc de chercher d'éventuelles fonctions paramétriques de $\lambda_0(t)$. Il a été montré par COX [8] et TSIATIS [6] que les estimations de β, fondées sur la vraisemblance partielle, ont des propriétés asymptotiquement voisines de celles fondées sur la vraisemblance habituelle. D'ailleurs l'efficacité de $\hat{\beta}$ par rapport aux autres estimateurs utilisant toute la vraisemblance au lieu d'une partie seulement, est très bonne dans la plupart des cas [10, 11, 12].

Comme estimateur de $\Lambda_0(t)$ où

$$\Lambda_0(t) = \int_{-\infty}^{t} \lambda(u, o)\, du$$

Tsiatis a proposé

$$\hat{\Lambda}_0(t) = \Sigma_{i \epsilon D(t)}\ 1 / \Sigma_{R(t_i)} \exp(\hat{\beta}' x_i) \qquad (8)$$

où D(t) désigne l'ensemble des décès survenant avant t (t inclus). On remarque que $\sum_{i \epsilon D(t_k)} \psi(i)$ n'est autre que $\sum_{i=1}^{k} \psi(i)$ les deux expressions étant d'usage courant dans la littérature. Cet estimateur est le même que celui de BRESLOW [7] quand il n'y a pas d'exaequos. Pour la variance de $\hat{\Lambda}_0(t)$ il montre qu'asymptotiquement elle est égale à

$$n \{\Sigma_{i \epsilon D(t)}\ 1/(\Sigma_{R(t_i)} \exp(\hat{\beta}' x_i))^2\} + \hat{\Psi}'(t)\, \Omega^{-1}\, \hat{\Psi}(t) \qquad (9)$$

où $\Omega^{-1}(\beta)$ est la matrice de variances et covariances de $\hat{\beta}$,

$$\hat{\Psi}'(t) = (\hat{\Psi}_1(t), \hat{\Psi}_2(t), \ldots\ldots\ldots\ldots\ldots \hat{\Psi}_p(t))$$

et $$\hat{\Psi}_r(t) = \Sigma_{i \epsilon D(t)} [\Sigma_{j \epsilon R(t_i)}\ x_{jr} \exp(\hat{\beta}' x_j) / \Sigma_{R(t_i)} \exp(\hat{\beta}' x_j))^2]$$

3. GENERALISATION

On a présenté au paragraphe 2 un estimateur pour le risque instantané cumulé, $\Lambda_0(t)$, et précisé une expression appropriée pour la variance asymptotique. Ceci est valable lorsqu'il n'y a pas d'exaequos parmi les temps de survie.

Le but de ce paragraphe est de généraliser ces résultats sur trois axes ; en se servant de la δ-méthode, on calcule les intervalles de confiance pour les fonctions de survie estimées à partir de (5) ; on trouve le résultat approprié quand les covariables ne sont plus contraintes de prendre les valeurs zéro et finalement, en utilisant l'estimateur de BRESLOW [7] au lieu de celui de TSIATIS, on peut travailler avec des exaequos parmi les temps de survie.

L'estimation de la courbe de survie est donnée par :

$$\hat{S}(t_k, \hat{\beta}) = \exp\{-\hat{\Lambda}_0(t_k) \exp(\hat{\beta}' x)\}$$

avec

$$\hat{\Lambda}_0(t_k) = \sum_{j=1}^{k} m_j / \sum_{i \in R_j} \exp(\hat{\beta}' x_i)$$

où m_j désigne le nombre de morts exaequos en t_j et $R_j = R(t_j)$.

Les résultats de TSIATIS pour var $\hat{\Lambda}_0(t_k)$ se généralisent dans le cas d'exaequos et nous mènent à

$$\text{var } \hat{\Lambda}_0(t_k) \simeq \sum_{j=1}^{k} m_j \Big/ \left\{ \sum_{i \in R_j} \exp(\hat{\beta}' x_i) \right\}^2 + \sum_{j=1}^{k} \sum_{s=1}^{k} \sum_{\ell} \sum_{h} \frac{m_j m_s \sum_{i \in R_j} x_{i\ell} \exp(\hat{\beta}' x_i) \sum_{i \in R_s} x_{ih} \exp(\hat{\beta}' x_i) \operatorname{cov}(\hat{\beta}_\ell, \hat{\beta}_h)}{\left\{ \sum_{i \in R_j} \exp(\hat{\beta}' x_i) \right\}^2 \left\{ \sum_{i \in R_s} \exp(\hat{\beta}' x_i) \right\}^2}$$

une expression qui se ramène à une formule identique à celle donnée par l'équation (9) dans le cas où tous les m_j valent un.

Soit

$$\hat{Y}(t_k, \hat{\beta}) = \log(-\log \hat{S}(t_k, \hat{\beta}))$$

Les intervalles de confiance à 100 $(1 - \alpha)$ % pour $\hat{Y}(t_k, \hat{\beta})$ sont donnés par

$$\hat{Y}(t_k, \hat{\beta}) \pm z_{1-\alpha/2} \{\text{var } \hat{Y}(t_k, \hat{\beta})\}^{1/2}$$

où Z est une variable aléatoire telle que

$$\text{pr}(-z_{1-\alpha/2} \leqslant Z \leqslant z_{1-\alpha/2}) \simeq 1 - \alpha$$

Lorsque $\hat{Y}(t_k, \hat{\beta})$ suit une loi normale avec moyenne $\log(-\log S(t_k, \beta))$ et écart type $\{\text{var } \hat{Y}(t_k, \hat{\beta})\}^{1/2}$ alors Z suit une loi normale réduite. On pourrait s'attendre à ce que la loi de probabilité de $Y(t_k, \hat{\beta})$ tende vers une loi normale assez rapidement puisque cette variable est une combinaison linéaire de $\hat{\beta}$. Quoiqu'il en soit les résultats de COX [8] et TSIATIS [6] ainsi que ceux afférents à tout estimateur du maximum de vraisemblance garantissent la normalité asymptotique de $\hat{Y}(t_k, \hat{\beta})$. Par la suite les intervalles de confiance de $\hat{S}(t_k, \hat{\beta})$ à 100 $(1 - \alpha)$ % sont alors donnés par

$$\exp[-\exp[\hat{Y}(t_k, \hat{\beta}) \pm z_{1-\alpha/2} \{\text{var } \hat{Y}(t_k, \hat{\beta})\}^{1/2}]]$$

où (cf. § 4):

$$\text{var } \hat{Y}(t_k, \hat{\beta}) \simeq \hat{\Lambda}_0^{-2}(t_k) \text{ var } \{\hat{\Lambda}_0(t_k)\} + \sum_{j} \sum_{m} x_j \operatorname{cov}(\hat{\beta}_j, \hat{\beta}_m) \{x_m + 2 \hat{\Lambda}_0^{-1}(t_k) \omega_m\} \tag{10}$$

et

$$\omega_m = -\sum_{j=1}^{k} m_j \left\{ \sum_{i \in R_j} x_{im} \exp(\hat{\beta}' x_i) \right\} \left\{ \sum_{i \in R_j} \exp(\hat{\beta}' x_i) \right\}^{-2}$$

4. CALCUL DE LA VARIANCE $\hat{Y}(t_k, \hat{\beta})$

$$\hat{Y}(t_k, \hat{\beta}) = \log \hat{\Lambda}_0(t_k) + \hat{\beta}'x$$

On obtient aussitôt, en faisant un calcul approché de variance par développement limité,

$$\text{var}\,\hat{Y}(t_k, \hat{\beta}) \simeq \hat{\Lambda}_0^{-2}(t_k)\,\text{var}\{\hat{\Lambda}_0(t_k)\} + 2\,\hat{\Lambda}_0^{-1}(t_k) \sum_j x_j \,\text{cov}\{\hat{\Lambda}_0(t_k), \hat{\beta}_j\} + \sum_j \sum_m x_j x_m \,\text{cov}(\hat{\beta}_j, \hat{\beta}_m) \qquad (11)$$

On dispose de toute l'information necessaire afin de calculer l'équation (11) à l'exception de cov $\{\hat{\Lambda}_0(t_k), \hat{\beta}_j\}$

On rappelle que

$$\text{cov}\{\hat{\beta}_\ell, \hat{\Lambda}_0(t_k)\} = E\{\hat{\beta}_\ell \hat{\Lambda}_0(t_k)\} - E(\hat{\beta}_\ell)\,E\{\hat{\Lambda}_0(t_k)\}$$

$$\simeq E\{\hat{\beta}_\ell \hat{\Lambda}_0(t_k)\} - \beta_\ell \Lambda_0(t_k) - \frac{1}{2}\,\beta_\ell \sum_n \sum_m \frac{\partial^2 \hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_n \partial\hat{\beta}_m}\,\text{cov}(\hat{\beta}_n, \hat{\beta}_m)$$

$$E\{\hat{\beta}_\ell \hat{\Lambda}_0(t_k)\} \simeq \beta_\ell \Lambda_0(t_k) + \frac{1}{2} \sum_n \sum_m \frac{\partial^2 \phi_\ell}{\partial\hat{\beta}_n \partial\hat{\beta}_m}\,\text{cov}(\hat{\beta}_n, \hat{\beta}_m)$$

où $$\phi_\ell = \sum_{j=1}^{k} \hat{\beta}_\ell\, m_j \Big/ \sum_{i \in R_j} \exp(\hat{\beta}' x_i) = \hat{\beta}_\ell \Lambda_0(t_k)$$

donc

$$\text{cov}\{\hat{\beta}_\ell, \hat{\Lambda}_0(t_k)\} \simeq \frac{1}{2}\left\{\sum_n \sum_m \left[\frac{\partial^2 \phi_\ell}{\partial\hat{\beta}_n \partial\hat{\beta}_m} - \hat{\beta}_\ell \frac{\partial^2 \hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_n \partial\hat{\beta}_m}\right] \text{cov}(\hat{\beta}_n, \hat{\beta}_m)\right\}$$

où β_ℓ a été remplacé par $\hat{\beta}_\ell$.

Cette équation se simplifie à

$$\sum_m \left(\frac{1}{2}\right)^{z(m)} \left[\frac{\partial^2 \phi_\ell}{\partial\hat{\beta}_\ell \partial\hat{\beta}_m} - \hat{\beta}_\ell \frac{\partial^2 \hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_\ell \partial\hat{\beta}_m}\right] \text{cov}(\hat{\beta}_\ell, \hat{\beta}_m)$$

où

z (m) est égal à 1 quand m est égal à ℓ et à zéro dans le cas contraire.

Après quelques étapes simples on montre que

$$\frac{\partial^2 \phi_\ell}{\partial\hat{\beta}_\ell \partial\hat{\beta}_m} - \hat{\beta}_\ell \frac{\partial^2 \hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_\ell \partial\hat{\beta}_m} = \begin{cases} \dfrac{2\,\partial\hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_\ell} & \text{si} \quad m = \ell \\ \dfrac{\partial\hat{\Lambda}_0(t_k)}{\partial\hat{\beta}_m} & \text{si} \quad m \neq \ell \end{cases}$$

d'où $$\text{cov}\{\hat{\beta}_\ell, \hat{\Lambda}_0(t_k)\} = \sum_m \omega_m \,\text{cov}(\hat{\beta}_\ell, \hat{\beta}_m)$$

avec $$\omega_m = \partial\hat{\Lambda}_0(t_k)/\partial\hat{\beta}_m$$

En employant cette expression avec l'équation (11) et quelques remaniements on retrouve l'équation (10) qui donne l'expression de la variance de $\hat{Y}(t_k, \hat{\beta})$.

5. UN EXEMPLE

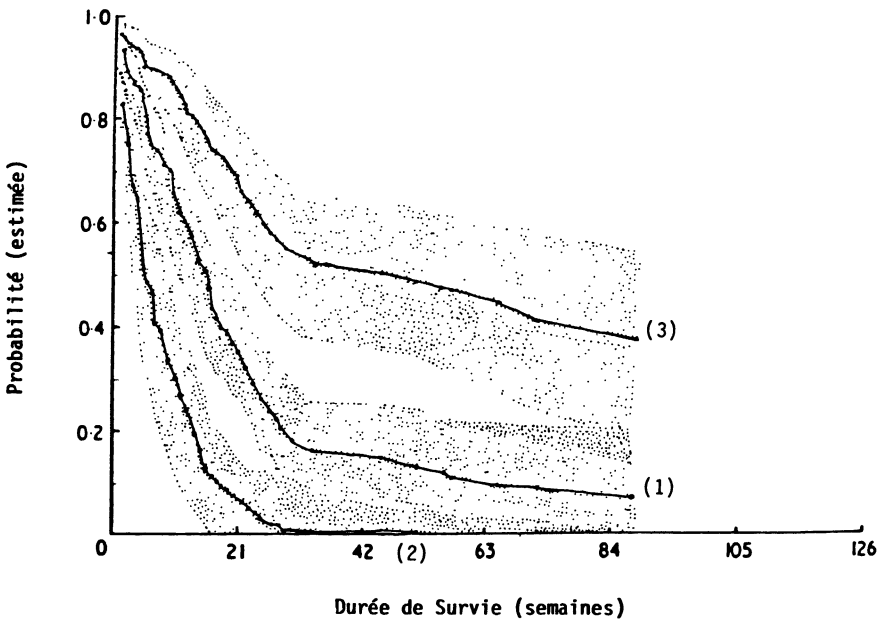
Dans un récent article [13] il a été montré que les niveaux des deux protéines étudiées dans le sang ont une importance pronostique chez les malades ayant un cancer de l'estomac. Pour utiliser la méthode de KAPLAN-MEIER il fallait remplacer les niveaux par des variables discrètes et l'échantillon a donc été partagé en trois sous-groupes à partir desquels ont été calculées des courbes de survie.

Le premier sous-groupe était composé des cas où les deux protéines dépassaient un niveau considéré comme seuil [13]. Pour le deuxième groupe une des deux protéines avait un niveau supérieur au seuil et pour le troisième ni l'une ni l'autre n'avait un niveau élevé.

En ce qui concerne une analyse fondée sur le modèle de COX [1] il n'était plus nécessaire de remplacer les variables quantitatives par des variables discrètes. On s'est donc servi des variables continues pour ne pas perdre une part importante d'information. Bien entendu, il fallait trouver les transformations appropriées pour que le modèle s'adapte, d'une façon satisfaisante aux données.

Pour illustrer les résultats on a calculé les estimations des courbes de survie et les intervalles de confiance dans trois cas; (1) les deux protéines prennent leurs valeurs moyennes (pour une des protéines on prend la valeur géométrique) (2) elles prennent les valeurs moyennes moins un écart type et (3) elles prennent les valeurs moyennes plus un écart type. La figure 1 illustre les trois cas.

Figure 1



(1) Valeurs moyennes.
(2) Valeurs moyennes plus un écart type.
(3) Valeurs moyennes moins un écart type.

REFERENCES BIBLIOGRAPHIQUES

[1] D.R. COX. –Regression models and life tables (with discussion). *J.R. Stat. Soc. B*, 34 : 187-220, 1972.

[2] R. KAY. – Proportional hazard regression models and the analysis of censored survival data. *Appl. Stat.* 26 : 227-237, 1977.

[3] D.M. STABLEIN, W.H. CARTER and G.L. WAMPLER. – Survival analysis of drug combinations using a hazards model with time-dependent covariates. *Biometrics*, 36 : 537-546, 1980.

[4] D. SCHOENFELD. – Chi-squared goodness of fit tests for the proportional hazards regression model. *Biometrika*, 67 : 145-153, 1980.

[5] P.K. ANDERSEN. – Testing goodness of fit of Cox's regression and life model. *Biometrics*, 38 : 67-77, 1982.

[6] A.A. TSIATIS. – A large sample study of Cox's regression model. *The Annals of Statistics*, 9 (1): 93-108, 1981.

[7] N.E. BRESLOW. – Covariance analysis of censored survival data. Biometrics, 30 : 89, 1974.

[8] D.R. COX. – Partial likelihood. *Biometrika*, 62 : 269-276, 1975.

[9] R.L. PRENTICE and J.D. KALBFLEISCH. – Hazard rate models with covariates. *Biometrics*, 35 : 25-39, 1979.

[10] J.D. KALBFLEISCH. – Some efficiency calculations for survival distributions. *Biometrika*, 61 : 31-38, 1974.

[11] B. EFRON. – The efficiency of Cox's likelihood function for censored data. *J. Am. Stat. Assoc.*, 72 : 555-565, 1977.

[12] R. KAY. – Some further asymptotic efficiency calculations for survival data regression models. *Biometrika*, 66 : 91-96, 1979.

[13] S.A. RASHID, J.O'QUIGLEY, A.T.R. AXON, E.H. COOPER. – Plasma protein profiles and prognosis in gastric cancer, *Br. J. Cancer*, 45 : 390-394, 1982.