Computer Methods and Programs in Biomedicine 63 (2000) 71–76

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model

Harald Heinzl *

Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

Received 30 August 1999; received in revised form 28 January 2000; accepted 1 February 2000

Abstract

Kent and O'Quigley (1988) apply the concept of information gain to define a measure of dependence (R-squared measure) between explanatory variables and a censored response variable within the framework of the Cox model. Two SAS macros to calculate this measure are presented. The first one is based on a Newton–Raphson search and makes use of the SAS IML procedure. The second one is a simple grid search using SAS DATA steps and Base-SAS procedures.

Keywords: Measure of dependence; Information gain; Explained variation; Cox model; Censored survival times; SAS macro

1. Introduction

The proportional hazards regression model of Cox [1] is the most popular model for analysing censored survival data in medical research. Consider a continuous survival time variable T. In a Cox regression model the conditional hazards function of T is assumed to be proportional to a function of the covariates X,

$$h(t|x) = h_0(t)\exp(x\beta)$$

where the baseline hazards function $h_0(t)$ is completely unspecified. Throughout the paper, covariates X will be represented by a p-dimensional row vector, and regression coefficients β will be represented by a p-dimensional column vector, respectively. The problem of interest is how to measure the dependence between T and X within the framework of the Cox model in a similar way as the coefficient of determination, R^2, does in the linear model. Due to censoring and the unspecified hazard function this is not a trivial task.

Various measures of dependence (R^2 measures, measures of explained variation) for the Cox model have been proposed in the literature, see for instance the survey of Schemper and Stare [7]. There is no doubt that the approach of Kent and O'Quigley [2] has resulted in a recommendable dependence measure with desirable statistical properties. Their approach has one serious handicap, which prevents its frequent use in medical research — because there is no explicit formula available, numerical optimization methods have

* Tel.: +43-1-404006686; fax: +43-1-404006687.

E-mail address: harald.heinzl@akh-wien.ac.at (H. Heinzl)

0169-2607/00/$ - see front matter
PII: S0169-2607(00)00073-0

to be employed. The following two SAS macros (SAS Institute, Cary, NC) are means to overcome this problem by providing easy-to-use tools within a popular and widely used statistical software package.

In Section 2 we describe theoretically how this measure of dependence is constructed and estimated. In Section 3 we describe the SAS programs to calculate this measure in practice after a Cox model has been fitted with PROC PHREG. The first one is based on PROC IML statements embedded in a SAS macro. It uses a Newton–Raphson search to find a numerical solution. The second program is for SAS users without access to the IML procedure. It uses SAS DATA steps and Base-SAS procedures to perform a simple grid search. In Section 4 both a worked example is given and the programs availability is described.

2. Computational methods and theory

We consider the linear regression model

$$Y = -\frac{\mu}{\alpha} - \frac{X\beta}{\alpha} + \frac{\varepsilon}{\alpha}$$

with outcome variable Y, covariate vector X, error term ε, and the model parameters $\theta = (\beta^T, \mu, \alpha)^T$, where β denotes the vector of regression coefficient for the covariates, μ denotes the regression coefficient for the constant term, and α denotes the scale parameter, $\alpha > 0$. This valid form of a linear regression model definition may be rather unfamiliar to the reader, however, it has been chosen, since it is more useful for technical reasons. The error term ε has a specified probability density function $f(.)$, and ε is independent of X. For instance, ε could have a standard normal distribution, $\varepsilon \sim N(0, 1)$. Of course, if we were solely interested in the normal error model, we would use the more common notation $\sigma = \alpha^{-1}$.

The conditional distribution of Y given X has probability density function $f(y \mid x;\ \theta)$, which is related to the error density by $f(y \mid x;\ \theta) = \alpha f(\alpha y + \mu + x\beta)$. Let $G(\mathrm{d}x)$ denote the marginal distribution of X, let 0_p denote the p-dimensional null vector, and let $\theta_1 = (\beta_1^T, \mu_1, \alpha_1)^T$ denote the true values of the model parameters, generally with $\beta_1 \neq 0_p$. The expected log likelihood of θ under θ_1 is defined by

$$\Phi(\theta;\ \theta_1) = \iint \log\{f(y|x;\ \theta)\} f(y|x;\ \theta_1)\ \mathrm{d}y G(\mathrm{d}x)$$

Consider the hypotheses H_0: '$\beta = 0_p$' and H_1: 'no restrictions on β'. Let θ_0 denote the θ maximizing $\Phi(\theta;\ \theta_1)$ over all θ satisfying H_0. That is, because $\theta_0 = (0_p^T, \mu_0, \alpha_0)^{\mathrm{T}}$, just the appropriate values for μ_0 and $\alpha_0 > 0$ have to be computed. It is obvious, that the vector of the true model parameter values θ_1 is the θ maximizing $\Phi(\theta;\ \theta_1)$ over all θ satisfying H_1.

A measure of dependence between Y and X can be determined by measuring and appropriately scaling the 'distance' between H_1 and H_0. By using twice the Kullback and Leibler [3] information gain to measure the distance between H_1 and H_0,

$$\Gamma = \Gamma(\mathrm{H}_1{:}\mathrm{H}_0;\ \theta_1, G) = 2\{\Phi(\theta_1;\ \theta_1) - \Phi(\theta_0;\ \theta_1)\}$$

Kent [4] proposed as measure of dependence

$$\rho_{\mathrm{IG}}^2 = 1 - \exp(-\Gamma)$$

Note that ρ_{IG}^2 depends on both the true parameter θ_1 of the conditional distribution of Y given X and the marginal distribution of X. The subscript 'IG' stands for the construction principle of ρ_{IG}^2, which is based on information gain.

Applying ρ_{IG}^2 to a normally distributed error density results in the ordinary coefficient of determination R^2. This fact leads to the question, whether the construction principle of ρ_{IG}^2 can be employed to measure dependence within the framework of Cox's regression model, also? The problem with Cox's model is due to its unknown error distribution, because the baseline hazards function $h_0(t)$ is completely unspecified. That is, the conditional distribution of T given X is specified only up to a monotone transformation of T, so that for any strictly monotone increasing function $\phi(.)$, $T^* = \phi(T)$ gives the same Cox regression coefficients as T. This fact is utilized by Kent and O'Quigley's approach [2]. By choosing ϕ in an appropriate way it can be ensured that the baseline hazards function is proportional to a power of t, $h_0^*(t) = \alpha \exp(\mu) t^{\alpha - 1}$ for any choice of

μ and $\alpha > 0$. Thus, the conditional distribution T^* given X follows a Weibull distribution, and $Y^* = \log(T^*)$ follows a linear regression model

$$Y^* = -\frac{\mu}{\alpha} - \frac{X\beta}{\alpha} + \frac{\varepsilon^*}{\alpha}$$

where the error term ε^* follows a standard extreme value distribution with density $f(z) = \exp\{z - \exp(z)\}$ and variance $\psi'(1) = \pi^2/6$. Here $\psi(.)$ and $\psi'(.)$ stand for the digamma function and its derivative, respectively. One unusual part of the linear model definition becomes clear now: Large values of $X\beta$ in the Cox model stand for high risk patients with small expected survival times, and vice versa. Therefore $-X\beta$ is used in the corresponding Weibull model (extreme value model).

Kent and O'Quigley's idea [2] can be put in other words: because any 'squeezing' or 'stretching' of the time axis does not change the results of the semiparametric Cox model, it should not change the results of a measure of dependence based on Cox model, also. Each Cox model result can be seen as a member of a class of equivalent results. Measuring the dependence between survival time and covariates within this class should yield identical values. That is, the dependence between T and X, $\phi_1(T)$ and X, $\phi_2(T)$ and X, $\phi_3(T)$ and X,..., should be the same as long as $\phi_1(.)$, $\phi_2(.)$, $\phi_3(.)$,... are strictly monotone transformations (i.e. they just 'squeeze' and 'stretch' the time axis). Now, pick a suitable representative out of this class of equivalent Cox model results, and apply the construction principle for the dependence measure based on information gain ρ^2_{IG}.

Kent and O'Quigley [2] denoted their measure of dependence for Cox model by ρ^2_{W} to emphasize the relationship to the Weibull distribution. Given x, the expected log likelihood takes the form

$$\begin{aligned}\Phi(\theta; \theta_1, x) &= \int_{-\infty}^{+\infty} \log\{\alpha f(\alpha y + \mu + x\beta)\}\alpha_1 f(\alpha_1 y + \mu_1 + x\beta_1) \\ &\times \mathrm{d}y = \log(\alpha) + \frac{\alpha}{\alpha_1}\gamma'(1) + b - \exp(b)\gamma\left(\frac{\alpha}{\alpha_1} + 1\right)\end{aligned}$$

where $b = \mu + x\beta - (\alpha/\alpha_1)(\mu_1 + x\beta_1)$. Here $\gamma(.)$ and $\gamma'(.)$ denote the gamma function and its derivative, respectively. This nonstandard notation is chosen to avoid confusion with the symbol Γ, which is already used for denoting information gain. The constant $\gamma'(1) = -0.577215... = \psi(1)$ is the negative value of Euler's constant.

It can be shown, that ρ^2_{W} does not depend on the choice of μ_1 and $\alpha_1 > 0$, they can be given arbitrary values. A natural choice is $\mu_1 \equiv 0$ and $\alpha_1 \equiv 1$, which corresponds to a constant baseline hazards function equal to one for H_1.

Now we can estimate the measure of dependence between T and X, $\tilde{\rho}^2_{\mathrm{W}} = 1 - \exp(-\tilde{\Gamma})$. Assume that for a study with n patients censored survival data have been observed with survival times t_i, censoring indicators c_i, and p-dimensional covariate vectors x_i, $i = 1... n$. Fitting a Cox regression model under H_1 to the data, that is, using all p covariates, yields estimated vector $\tilde{\beta}$ of regression coefficients. For calculating the estimator $\tilde{\rho}^2_{\mathrm{W}}$ we consider $\tilde{\theta}_1 = (\tilde{\beta}^T, 0, 1)^T$ as true parameter values. Thus, $\tilde{\Gamma} = \Gamma\{\mathrm{H}_1{:}\mathrm{H}_0; \tilde{\theta}_1, G_n(\mathrm{d}x)\}$, where $G_n(\mathrm{d}x)$ denotes the empirical distribution of X. The main problem is to find an estimate for θ_0, $\tilde{\theta}_0 = (0_p^T, \tilde{\mu}_0, \tilde{\alpha}_0)^T$. The empirical expected log likelihood

$$\Phi(\theta; \tilde{\theta}_1) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\theta; \tilde{\theta}_1, x_i)$$

has to be numerically maximized with respect to μ and α, $\alpha > 0$. Taking partial derivatives and setting them to zero finally yields an explicit solution for $\tilde{\mu}_0$,

$$\tilde{\mu}_0 = -\log(\gamma(\tilde{\alpha}_0 + 1)) - \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp(-\tilde{\alpha}_0 x_i \tilde{\beta})\right)$$

and an implicit solution for $\tilde{\alpha}_0$,

$$\xi(\alpha) := \psi(1) - \psi(\alpha) + \sum_{i=1}^{n} \frac{\exp(-\alpha z_i)}{\sum_{j=1}^{n} \exp(-\alpha z_j)} z_i = 0$$

where $z_i = x_i\tilde{\beta} - \bar{x}\tilde{\beta}$, $i = 1... n$, and the vector $\bar{x}$ contains the mean values of the p covariates. A lot of numerical techniques are available to solve the nonlinear equation $\xi(\alpha) = 0$. We consider two, (a) the Newton–Raphson method, and (b) a simple grid search. The former is chosen for its speed of convergence, the latter for its simplicity.

The Newton–Raphson method [5] is an iterative algorithm, which takes the form $\alpha^{(k+1)} = \alpha^{(k)} - \xi(\alpha^{(k)})/\xi'(\alpha^{(k)})$. Both an initial guess $\alpha^{(0)}$ and a stopping criterion are required. Here we use $\alpha^{(0)} = 1$ and $|\xi(\alpha^{(k)})| \leq 10^{-6}$, respectively. To save potential numerical trouble caused by the condition $\alpha > 0$, $\exp(\tau)$ can be substituted for α.

The grid search method makes use of the fact that $\tilde{\alpha}_0 \in (0, \tilde{\alpha}_1] \equiv (0, 1]$, see Appendix A for a proof. Note the correspondence to the normal error model, where it is well-known that the variance of the simple constant model, σ_0^2, is always greater or equal to the variance of the model with covariates, σ_1^2, and $\sigma = \alpha^{-1}$. First, lay a grid $\{k/g: k = 1\ldots g\}$ of equally spaced evaluation points over the interval (0,1], where the integer g determinates how close-meshed this grid should be. Secondly, evaluate either $\xi(\alpha)$ or $\Phi(\theta; \tilde{\theta}_1)$, $\theta = (0_p^T, \tilde{\mu}_0, \alpha)^T$, for all values of the grid, that is $\alpha = 1/g, 2/g, \ldots, 1$. Finally, set $\tilde{\alpha}_0$ to the value of the grid, which either minimizes $|\xi(\alpha)|$ or, equivalently, maximizes the empirical expected log likelihood with respect to H_0.

After we have found a numerical solution for $\tilde{\alpha}_0$, we can compute $\tilde{\rho}_W^2 = 1 - \exp(-\tilde{\Gamma})$, where

$$\tilde{\Gamma} = 2\left[(1 - \tilde{\alpha}_0)\psi(1) + \log\{\gamma(\tilde{\alpha}_0)\} + \log\left\{\frac{1}{n}\sum_{i=1}^{n}\exp(-\tilde{\alpha}_0 z_i)\right\}\right]$$

We can explore the formula for two important borderline situations. For $\tilde{\alpha}_0 = 1$, the estimated information gain $\tilde{\Gamma}$ between H_1 and H_0 will equal zero, and $\tilde{\rho}_W^2 = 0$. This can only happen, if the covariates do not influence survival times. More formally this corresponds to $\tilde{\beta} = 0_p$, see also Appendix A. On the other hand, if the covariates influence on survival times is enormously strong, $\tilde{\alpha}_0$ will tend to a value close to zero, the gamma function $\gamma(\tilde{\alpha}_0)$ will tend to infinity, $\tilde{\Gamma}$ will tend to infinity, too, and finally, $\tilde{\rho}_W^2$ will tend to one as is to be expected.

3. Program description

The SAS macros are called KENTOQNR and KENTOQGS. Here 'KENTOQ' stands for Kent and O'Quigley, who invented the measure of dependence in question. 'NR' and 'GS' stand for the specific numerical procedure in use, that is Newton–Raphson or grid search, respectively. Several simple macro statements allow the user to specify the Cox model, for which an estimate for the measure of dependence should be computed. Both macros share most of the statements, different statements only occur with respect to the numerical procedures. The meaning of the various common statements is as follows:

- TITLE: title of the analysis
- DATA: SAS data set which contains the survival data, thereby the name '_xbeta' should be avoided, because it is reserved for internal use only
- TIME: survival time variable
- STATUS: status variable
- CENSLIST: value(s) of status variable which indicate censoring of survival times, the default value is 0
- COV: names of covariates

The next two statements are solely for KENTOQNR-macro, they are meant for controlling the Newton–Raphson iterations:

- MAXITER: specifies the maximum number of Newton–Raphson steps allowed, the default value is 25
- CONVCRIT: convergence criterion for Newton–Raphson method, the iteration sequence stops when $|\xi(\alpha^{(k)})| \leq$ CONVCRIT, with 10^{-6} set as the default value

The next statement is solely for KENTOQGS-macro, it is meant for controlling the grid search algorithm:

- GRID: integer value g determinating the number of equally spaced evaluation points in the grid, the default value is set to 100

Submitting the appropriate macro statements will compute the requested Cox model and $\tilde{\rho}_W^2$, labelled by 'rho-squared W' in the KENTOQNR-macro output or 'RHO2_W' in the KENTOQGS-macro output, respectively. From the authors experience, the Newton–Raphson method usually will converge in two to four steps for the default convergence criterion value 10^{-6}. The grid search method will give excellent approximations for the default value 100 of the GRID-statement. Of course, if PROC IML is available, the macro KENTOQNR will be first choice. However, SAS users with no access to PROC IML will have a reasonable alternative in the macro KENTOQGS, which is based on brute force of the computer.

4. Sample of typical program run and program availability

The programs will be illustrated by using the Veteran's Administration lung cancer data published in Kalbfleisch and Prentice [6]. This data set consists of survival times, censoring status and various covariates of 137 males with inoperable lung cancer. Schemper and Stare [7] report the results of a Cox model fit with covariates treatment (dichotomous), age in years (continuous), histology (qualitative with four levels) and Karnofsky performance index (continuous). They also report an estimated value for the measure of dependence, $\tilde{\rho}^2_{\mathrm{W}} = 0.39$. That is, these four covariates explain 39% of the variability in the survival time.

The analysis of Schemper and Stare [7] can be repeated now with either of the SAS macros KENTOQNR and KENTORGS. Assume that the data are stored in the temporary SAS data set VALUNGCA, and that the covariate histology is represented by three dummy variables, H2, H3, and H4. The variable names of the other covariates are TREATM, AGE and PERF. By relying upon the default settings for the statements CENSLIST, MAXITER and CONVCRIT, the call of the KENTOQNR-macro is as simple as follows.

```
%KENTOQNR
(TITLE=Lung cancer data of
Kalbfleisch and Prentice,
DATA=valungca, TIME=time, STATUS=
status,
COV=treatm age h2 h3 h4 perf);
```

The output consists of the Cox model fit from PROC PHREG (omitted) and an addendum reporting $\tilde{\rho}^2_{\mathrm{W}}$.

```
Lung cancer data of Kalbfleisch and
Prentice
Measure of dependence of Kent and
O'Quigley
(computation based on Newton-Raph-
son method)
rho-squared W: 0.3858
```

By relying upon the default settings for the statements CENSLIST and GRID, the call of the KENTOQGS-macro is equivalent to that of the KENTOQNR-macro shown above, so it will be omitted here. The output is similar, too. It consists of the Cox model fit (omitted) and the value of $\tilde{\rho}^2_{\mathrm{W}}$. Note the close agreement between Newton–Raphson and grid search result, this is quite common.

```
Lung cancer data of Kalbfleisch and
Prentice
Measure of dependence of Kent and
O'Quigley
computed by a grid search with 100
steps
  RHO2_W
  0.3858
```

The SAS macros and the lung cancer data set are available via world wide web at 'http://www.akh-wien.ac.at/Harald.Heinzl/sasmacro/kentoq.zip'.

Appendix A

We want to show that the solution $\tilde{\alpha}_0$ of the nonlinear equation $\xi(\alpha) = 0$ of Section 2 is constrained to the interval (0,1]. Define $\xi(\alpha) = \xi_1(\alpha) + \xi_2(\alpha)$, where $\xi_1(\alpha) = \psi(1) - \psi(\alpha)$ and

$$\xi_2(\alpha) = \sum_{i=1}^{n} \frac{\exp(-\alpha z_i)}{\sum_{j=1}^{n} \exp(-\alpha z_j)} z_i$$

Because the digamma function $\psi(\alpha)$ is continuous and strictly monotonically increasing for $\alpha > 0$, $\xi_1(\alpha)$ is continuous and strictly monotonically decreasing for $\alpha > 0$, $\xi_1(\alpha) > 0$ for $\alpha \in (0, 1)$, and $\xi_1(1) = 0$.

The function $\xi_2(\alpha)$ is continuous and monotonically decreasing, $\xi_2(0) = 0$, and $\xi_2(\alpha) \leq 0$ for $\alpha > 0$. This is due to the fact that $\xi_2(\alpha)$ is the weighted mean of $z_1 \ldots z_n$, where the weights are $w_i = \exp(-\alpha z_i)/\Sigma_{j=1}^{n} \exp(-\alpha z_j)$, $i = 1 \ldots n$, and $w_1 +$

$\ldots + w_n = 1$. For $\alpha = 0$ the weights become $w_i = 1/n$, and the weighted mean reduces to the common mean, $\bar{z} = 0$. For $\alpha > 0$ the weights for smaller observations of $z_1 \ldots z_n$ become larger, and vice versa. Finally, for very large values of α, the weight of the smallest observation will tend to a value of one and all the other weights will vanish, that is,

$$\lim_{\alpha \to \infty} \{\xi_2(\alpha)\} = \min_{i=1\ldots n} \{z_i\} \leq 0$$

Summing up, since $\xi_1(\alpha)$ is a continuous and strictly monotonically decreasing function for $\alpha > 0$, which crosses zero at $\alpha = 1$, and because $\xi_2(\alpha)$ is a continuous and monotonically decreasing function, which is zero at $\alpha = 0$, the unique solution $\tilde{\alpha}_0$ of the equation $\xi_1(\alpha) + \xi_2(\alpha) = 0$ has to be in interval (0,1]. Note that $\tilde{\alpha}_0 = 1$ will only show up when $\xi_2(1) = 0$. That is, $z_1 = \ldots = z_n = 0$, which is either due to collinearity of the covariates or $\tilde{\beta} = 0_p$. Of course, the former is not relevant here, since the Cox model would not be estimable then. However, $\tilde{\beta} = 0_p$ represents an important special case, since it corresponds to the situation where the covariates are of no influence on survival times. On the other hand, the stronger the covariates will influence the survival times, the more $z_1 \ldots z_n$ will spread out, the steeper $\xi_2(\alpha)$ will decrease, and the closer to zero the value of $\tilde{\alpha}_0$ will be.

References

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

[2] J.T. Kent, J. O'Quigley, Measures of dependence for censored survival data, Biometrika 75 (1988) 525–534.

[3] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86.

[4] J.T. Kent, Information gain and a general measure of correlation, Biometrika 70 (1983) 163–173.

[5] G.W. Reddien, Newton–Raphson methods, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, vol. 6, Wiley, New York, 1985, pp. 210–212.

[6] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, Wiley, New York, 1980.

[7] M. Schemper, J. Stare, Explained variation in survival analysis, Stat. Med. 15 (1996) 1999–2012.