Biometrical Journal **44** (2002) 6, 671–682

A Measure of Dependence for the Stratified Cox Proportional Hazards Regression Model

HARALD HEINZL[1, *], JANEZ STARE[2] and MARTINA MITTLBÖCK[1]

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

[2] Department of Biomedical Informatics, University of Ljubljana, Vrazov Trg 2, SLO-1105 Ljubljana, Slovenia

Abstract

KENT and O'QUIGLEY (1988) apply the concept of information gain to measure both global and partial dependence between explanatory variables and a censored response within the framework of the proportional hazards regression model of COX (1972). The definition of this measure is extended to cover also the stratified Cox model.

Key words: Information gain; Kullback-Leibler distance; Weibull distribution; Survival analysis; Censored failure time data; Semi-parametric model; R-squared measure.

1. Introduction

Consider a continuous survival time variable T and covariate vectors $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$, which are row vectors of dimension $p^{(1)}$, $p^{(2)}$ and $p^{(3)}$, respectively. To assess the effects of $\mathbf{X}^{(1)}$ on T, thereby considering $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ as confounding variables, we want to apply the proportional hazards regression model of COX (1972), which is the most popular model for analysing survival data in medical research. Now assume that the proportional hazards assumption does not hold with respect to covariates $\mathbf{X}^{(3)}$. With problems of this kind it is often possible to split the $\mathbf{X}^{(3)}$ variables into homogeneous subgroups (strata), represented by the m-level-factor S. For instance, $\mathbf{X}^{(3)}$ could consist of sex of the patient and some hormone status variables, and for some medical reasons a 3-level-factor S could appear appropriate with levels {male, premenopausal female, postmenopausal female}. Now a stratified Cox proportional hazards regression model can be defined,

$$h_s(t \mid \mathbf{x}) = h_{0s}(t) \exp\left(\mathbf{x}^{(1)}\boldsymbol{\beta}^{(1)} + \mathbf{x}^{(2)}\boldsymbol{\beta}^{(2)}\right), \tag{1.1}$$

* Corresponding author: harald.heinzl@univie.ac.at

 0323-3847/02/0609-0671 $ 17.50+.50/0

$s = 1, \ldots, m$, where the baseline hazards functions $h_{0s}(t)$ are completely unspecified and mutually unrelated. Throughout the paper, covariate vectors will be considered as row vectors, and regression coefficient vectors will be considered as column vectors, respectively. A crucial feature of model (1.1) is its invariance under differentiable, strictly monotonic increasing transformations acting on the time scale in each stratum (KALBFLEISCH and PRENTICE, 1980, p. 88).

The aim of this paper is to propose a measure of partial dependence for model (1.1). That is, the dependence between T and $\mathbf{X}^{(1)}$ is measured, after the effects of confounders $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ have been accounted for by regression and stratification, respectively.

If $\mathbf{X}^{(3)}$ is absent, $p^{(3)} = 0$. This is the so called unstratified case, since only one stratum exists, and $m = 1$. Various dependence measures (R-squared measures) have been proposed for the unstratified Cox model (SCHEMPER and STARE, 1996). Discussion on which measure to use in practice is still ongoing, since the semiparametric nature of the Cox model allows more than one sensible generalization of the definition of linear model R^2, where our notion of *measured dependence* (*explained randomness, explained variation*) usually is derived from. However, there is no doubt that the approach of KENT and O'QUIGLEY (1988) has resulted in a dependence measure with desirable statistical properties. The Kent and O'Quigley measure is based on the methodological considerations of KENT (1983), who used the concept of information gain (KULLBACK and LEIBLER, 1951) to generalise the usual squared multiple correlation coefficient (linear model R^2) to a wider class of parametric models of dependence. The idea was previously also used by LINFOOT (1957) to define a generalised joint correlation coefficient.

In Section 2 we give a brief outline how to construct the KENT and O'QUIGLEY measure (1988) for the unstratified Cox model. The measure will be generalized for the stratified Cox model in Section 3. If both the stratified measure and the unstratified partial measure are applicable, then their results will be expected to be similar. This will be explored in Section 4 by means of a simulation study. The stratified measure will be applied to the Veteran's Administration lung cancer data published in KALBFLEISCH and PRENTICE (1980) in Section 5. Finally, a discussion is given in Section 6.

2. Background: The Unstratified Case

This Section provides a brief description of the approach of KENT and O'QUIGLEY (1988) for the unstratified Cox model, that is, $p^{(3)} = 0$ and $m = 1$. For the sake of redundancy the strata indicator has been omitted throughout this Section. We consider the linear model

$$y = -\frac{\mu}{\alpha} - \frac{\mathbf{x}\boldsymbol{\beta}}{\alpha} + \frac{\varepsilon}{\alpha} = -\frac{\mu}{\alpha} - \frac{\mathbf{x}^{(1)}\boldsymbol{\beta}^{(1)}}{\alpha} - \frac{\mathbf{x}^{(2)}\boldsymbol{\beta}^{(2)}}{\alpha} + \frac{\varepsilon}{\alpha}\,, \tag{2.1}$$

where the error term ε has a specified probability density function $f(z)$ say, and ε is independent of the covariates $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$. Let $G(\mathrm{d}\mathbf{x})$ denote the marginal distribution of $\mathbf{X}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{(1)T}, \boldsymbol{\beta}^{(2)T}, \mu, \alpha)^T$ denote the parameters of the model, $\alpha > 0$, and let $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{(1)T}, \boldsymbol{\beta}_1^{(2)T}, \mu_1, \alpha_1)^T$ denote the true values of the parameters, generally with $\boldsymbol{\beta}_1^{(1)} \neq \mathbf{0}_{p^{(1)}}$, where $\mathbf{0}_{p^{(1)}}$ is the $p^{(1)}$-dimensional zero vector. Consider the nested hypotheses $\mathrm{H}_0 : \boldsymbol{\beta}^{(1)} = \mathbf{0}_{p^{(1)}}$ and H_1: no restrictions on $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)T}, \boldsymbol{\beta}^{(2)T})^T$. KENT (1983) proposed

$$\varrho_{\mathrm{IG}}^2 = 1 - \exp\left(-\Gamma\right)$$

to measure the dependence between Y and $\mathbf{X}^{(1)}$ after allowing for the regression on $\mathbf{X}^{(2)}$. If $p^{(2)} = 0$, then $\mathbf{X}^{(2)}$ will be empty, and ϱ_{IG}^2 will be called a common R-squared measure. If $p^{(2)} > 0$, then ϱ_{IG}^2 will be called a partial R-squared measure. Γ is twice the KULLBACK and LEIBLER (1951) information gain,

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

which is used to measure the *distance* between H_1 and H_0. $\boldsymbol{\theta}_0$ is defined to be the value of $\boldsymbol{\theta}$ maximizing the expected log likelihood $\Phi(\boldsymbol{\theta};\ \boldsymbol{\theta}_1)$ over all $\boldsymbol{\theta}$ satisfying H_0, where

$$\Phi(\boldsymbol{\theta};\ \boldsymbol{\theta}_1) = \iint \log\{f(y \mid \mathbf{x};\ \boldsymbol{\theta})\}\, f(y \mid \mathbf{x};\ \boldsymbol{\theta}_1)\,\mathrm{d}y\, G(\mathrm{d}\mathbf{x})\,. \tag{2.2}$$

The probability density function of the conditional distribution of Y given $\mathbf{X}$ is denoted by $f(y \mid \mathbf{x};\ \boldsymbol{\theta})$. In the following we will relate it to the error density by $f(y \mid \mathbf{x};\ \boldsymbol{\theta}) = \alpha f(\alpha y + \mu + \mathbf{x}\boldsymbol{\beta})$. Note that Γ depends on H_1, H_0, the true parameter $\boldsymbol{\theta}_1$ and the marginal distribution of $\mathbf{X}$. We can transform ϱ_{IG}^2 into the general form of a dependence measure (proportion of *explained randomness*, proportion of *explained variation*),

$$\varrho_{\mathrm{IG}}^2 = 1 - \frac{\exp\{-2\Phi(\boldsymbol{\theta}_1;\ \boldsymbol{\theta}_1)\}}{\exp\{-2\Phi(\boldsymbol{\theta}_0;\ \boldsymbol{\theta}_1)\}} = 1 - \frac{M(Y \mid \mathbf{X})}{M(Y \mid \mathbf{X}^{(2)})}\,,$$

where $M(Y \mid \mathbf{X}^{(2)})$ and $M(Y \mid \mathbf{X})$ denote the *residual randomness* of Y under the hypotheses H_0 and H_1, respectively (KENT and O'QUIGLEY, 1988). That is, ϱ_{IG}^2 is the proportion of *residual randomness* unexplained under H_0, which can be explained through the inclusion of the covariates $\mathbf{X}^{(1)}$ under H_1.

Calculating ϱ_{IG}^2 for a normally distributed error density in model (2.1) yields the ordinary linear model R^2. However, it is not obvious at first sight how to employ the construction principle of ϱ_{IG}^2 to measure dependence within the framework of the unstratified Cox proportional hazards regression model. The problem with this model is due to its unknown error distribution, since the baseline hazards function $h_0(t)$ is unspecified. That is, the conditional distribution of T given $\mathbf{X}$ is specified only up to a strictly monotonic transformation of T, and $\phi(T)$ gives the same Cox regression coefficients as T for any strictly monotonic increasing function ϕ. In other words, covariates enter the model through their actual distributions whereas the survival time outcome is dealt with in a fully non-parametric fashion.

Here KENT and O'QUIGLEY (1988) set in by pointing out an interesting analogy to non-parametric correlation coefficients. For example, both the Spearman and Fisher-Yates correlations can be viewed as product-moment correlations between two variables after transforming them to have uniform and normal marginal distributions, respectively. Shifted to the problem at present this means that an appropriate transformation ϕ for the survival time outcome T given $\mathbf{X}$ has to be chosen first, before the construction principle of ϱ^2_{IG} can be applied to define a semi-parametric R^2-measure for the Cox model.

A simple and *natural* choice for a conditional survival distribution in the proportional hazards context seems to be the exponential distribution, or, more generally, its location-scale family, that is the Weibull distribution family. In fact, ϕ can be chosen for the conditional distribution of $T^* = \phi(T)$ to follow a Weibull distribution and the baseline hazards function is proportional to a power of t, $h^*_0(t) = \alpha \exp(\mu)\, t^{\alpha-1}$, for any choice of real numbers μ and $\alpha > 0$. If T^* given $\mathbf{X}$ follows a Weibull distribution, then $\log(T^*)$ will result in the linear regression model (2.1), and the error term will follow a standard extreme value distribution with density $f(z) = \exp\{z - \exp(z)\}$, see LAWLESS (1982). It remains to compute ϱ^2_{IG} for this type of error term distribution. Given $\mathbf{x}$, the expected log likelihood takes the form

$$\begin{aligned}\Phi(\boldsymbol{\theta};\, \boldsymbol{\theta}_1,\, \mathbf{x}) &= \int_{-\infty}^{+\infty} \log\{\alpha f(\alpha y + \mu + \mathbf{x}\boldsymbol{\beta})\}\, \alpha_1 f(\alpha_1 y + \mu_1 + \mathbf{x}\boldsymbol{\beta}_1)\, \mathrm{d}y \\ &= \log(\alpha) + \frac{\alpha}{\alpha_1}\, \gamma'(1) + b - \exp(b)\, \gamma\left(\frac{\alpha}{\alpha_1} + 1\right),\end{aligned}$$

where $b = \mu + \mathbf{x}\boldsymbol{\beta} - (\alpha/\alpha_1)\,(\mu_1 + \mathbf{x}\boldsymbol{\beta}_1)$, $\gamma(.)$ denotes the gamma function, and $-\gamma'(1) = 0.577\ldots$ is Euler's constant. KENT and O'QUIGLEY (1988) denoted the finally resulting dependence measure by ϱ^2_{W} in order to emphasize the relationship to the Weibull distribution.

An estimate for ϱ^2_{W} can be found by estimating Γ first so that $\tilde{\varrho}^2_W = 1 - \exp(-\tilde{\Gamma})$. Remember that the definition of Γ is based on the vector of true parameter values $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{(1)\,T},\, \boldsymbol{\beta}_1^{(2)\,T},\, \mu_1,\, \alpha_1)^T$. Since Γ does not depend on the choice of μ_1 and $\alpha_1 > 0$, the values $\mu_1 = 0$ and $\alpha_1 = 1$ are chosen for the sake of convenience. However, for $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_1^{(1)\,T},\, \boldsymbol{\beta}_1^{(2)\,T})^T$ we have to insert an appropriate estimate. Let us assume that censored survival data $\{(t_i,\, c_i,\, \mathbf{x}_i);\ i = 1, \ldots, n\}$ are available with survival times t_i, censoring indicators c_i, and $(p^{(1)} + p^{(2)})$-dimensional covariate vectors $\mathbf{x}_i$. Fitting a Cox regression model under H_1 to the data, that is by using all $p^{(1)} + p^{(2)}$ covariates, yields the vector $\tilde{\boldsymbol{\beta}}_{\mathrm{Cox},\,1}$ of estimated regression coefficients which is substituted for $\boldsymbol{\beta}_1$ to get

$\tilde{\boldsymbol\theta}_1 = (\tilde{\boldsymbol\beta}^{(1)\,T}_{\text{Cox},\,1}, \tilde{\boldsymbol\beta}^{(2)\,T}_{\text{Cox},\,1}, 0, 1)^T$. The estimate $\tilde{\boldsymbol\theta}_0 = (\mathbf{0}^T_{p^{(1)}}, \tilde{\boldsymbol\beta}^{(2)\,T}_0, \tilde{\mu}_0, \tilde{\alpha}_0)^T$ for $\boldsymbol\theta_0$ is found by numerically maximizing the empirical expected log likelihood $\Phi(\boldsymbol\theta; \tilde{\boldsymbol\theta}_1) = \dfrac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol\theta; \tilde{\boldsymbol\theta}_1, \mathbf{x}_i)$ over all $\boldsymbol\theta = (\mathbf{0}^T_{p^{(1)}}, \boldsymbol\beta^{(2)\,T}, \mu, \alpha)^T$ satisfying H_0, $\alpha > 0$. To save numerical trouble $\exp(\tau)$ can be substituted for α. Finally, $\tilde{\Gamma} = \Gamma(\mathrm{H}_1 : \mathrm{H}_0; \tilde{\boldsymbol\theta}_1, G_n(d\mathbf{x})) = 2\{\Phi(\tilde{\boldsymbol\theta}_1; \tilde{\boldsymbol\theta}_1) - \Phi(\tilde{\boldsymbol\theta}_0; \tilde{\boldsymbol\theta}_1)\}$, where $G_n(d\mathbf{x})$ denotes the empirical distribution of the covariates $\mathbf{X}$.

Due to the lack of an according closed-form solution for $\tilde{\varrho}^2_{\mathrm{W}}$, this measure has been considered difficult to compute (SCHEMPER and STARE, 1996). However, the fitting of a Cox model itself shares the same level of computing difficulty. In both cases, numerical optimisation techniques like the Newton-Raphson method (REDDIEN, 1985; HEINZL, 2000) have to be employed. Nevertheless, KENT and O'QUIGLEY (1988) also proposed a closed-form approximation for ϱ^2_{W} by considering the error term in model (2.1) to follow a standard normal distribution. By using a more common notation for this case, that is $\sigma^2 = \alpha^{-2}$, they ended up with $\varrho^2_{\mathrm{W,A}} = 1 - (1/\sigma^2_0)$, where $\sigma^2_0 = 1 + \boldsymbol\beta^{(1)\,T}_1 \boldsymbol\Omega^{(11.2)} \boldsymbol\beta^{(1)}_1$, $\boldsymbol\Omega^{(11.2)} = \boldsymbol\Omega^{(11)} - \boldsymbol\Omega^{(12)} \times \{\boldsymbol\Omega^{(22)}\}^{-1}\, \boldsymbol\Omega^{(21)}$, and $\boldsymbol\Omega$ is the covariance matrix of $\mathbf{X}$ partitioned in the usual way. Just as in the case of ϱ^2_{W} the choice of μ_1 and $\sigma^2_1 > 0$ is irrelevant for the value of $\varrho^2_{\mathrm{W,A}}$ so that $\mu_1 = 0$ and $\sigma^2_1 = 1$ have been chosen. To estimate $\varrho^2_{\mathrm{W,A}}$ we just replace $\boldsymbol\beta^{(1)}_1$ and $\boldsymbol\Omega$ by $\tilde{\boldsymbol\beta}^{(1)}_{\text{Cox},\,1}$ and $\tilde{\boldsymbol\Omega}$, respectively, where $\tilde{\boldsymbol\Omega}$ denotes the sample covariance matrix of $\mathbf{X}$.

3. The Stratified Case

The approach of Kent and O'Quigley (1988) is extended to cover the stratified Cox regression model (1.1) as well, so that we are able to quantify the partial dependence between T and $\mathbf{X}^{(1)}$ after the effects of confounders $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ have been accounted for by regression and stratification, respectively. Remember that the $\mathbf{X}^{(3)}$ variables are split into strata, which are represented by the m-level-factor S. At first we develop a general definition of a stratified dependence measure based on information gain. For this we consider the stratified linear regression model

$$y = -\frac{\mu_s}{\alpha_s} - \frac{\mathbf{x}^{(1)}\boldsymbol\beta^{(1)}}{\alpha_s} - \frac{\mathbf{x}^{(2)}\boldsymbol\beta^{(2)}}{\alpha_s} + \frac{\varepsilon}{\alpha_s}\,, \tag{3.1}$$

$s = 1, \ldots, m$, where the error term ε follows a specified probability density function $f(z)$ say, and ε does not depend on covariates $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ or strata S. The vector of regression coefficients $\boldsymbol\beta = (\boldsymbol\beta^{(1)\,T}, \boldsymbol\beta^{(2)\,T})^T$ is constant over strata, and by analogy to Section 2 we consider the nested hypotheses $\mathrm{H}_0 : \boldsymbol\beta^{(1)} = \mathbf{0}_{p^{(1)}}$ and H_1 : no restrictions on $\boldsymbol\beta$. Define $\boldsymbol\theta = (\boldsymbol\beta^T, \mu_1, \ldots, \mu_m, \alpha_1, \ldots, \alpha_m)^T$, where

$\alpha_s > 0$, $s = 1, \ldots, m$. The expected log likelihood is

$$\begin{aligned}\Phi(\boldsymbol{\theta};\ \boldsymbol{\theta}_1) &= \sum_{s=1}^{m} \omega_s \Phi(\boldsymbol{\theta};\ \boldsymbol{\theta}_1, s)\\ &= \sum_{s=1}^{m} \omega_s \int \Phi(\boldsymbol{\theta};\ \boldsymbol{\theta}_1,\ \mathbf{x},\ s)\ G(\mathrm{d}\mathbf{x} \mid s)\\ &= \sum_{s=1}^{m} \omega_s \iint \log\{f(y \mid \mathbf{x},\ s;\ \boldsymbol{\beta},\ \mu_s,\ \alpha_s)\}\\ &\quad \times f(y \mid \mathbf{x},\ s;\ \boldsymbol{\beta}_1,\ \mu_{1s},\ \alpha_{1s})\,\mathrm{d}y\, G(\mathrm{d}\mathbf{x} \mid s)\end{aligned} \tag{3.2}$$

where $f(y \mid \mathbf{x},\ s;\ \boldsymbol{\beta},\ \mu_s,\ \alpha_s) = \alpha_s f(\alpha_s y + \mu_s + \mathbf{x}\boldsymbol{\beta})$ denotes the probability density function of the conditional distribution of Y given $\mathbf{X}$ and S, $G(\mathrm{d}\mathbf{x} \mid s)$ denotes the conditional distribution of $\mathbf{X}$ given S, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$ denotes the probability distribution of S. The *distance* between H_1 and H_0 can be measured by twice the weighted average of the stratum-specific information gains,

$$\begin{aligned}\Gamma &= \Gamma(\mathrm{H}_1 : \mathrm{H}_0;\ \boldsymbol{\theta}_1,\ \mathbf{G}, \boldsymbol{\omega})\\ &= 2\{\Phi(\boldsymbol{\theta}_1;\ \boldsymbol{\theta}_1) - \Phi(\boldsymbol{\theta}_0;\ \boldsymbol{\theta}_1)\}\\ &= \sum_{s=1}^{m} \omega_s [2\{\Phi(\boldsymbol{\theta}_1;\ \boldsymbol{\theta}_1,\ s) - \Phi(\boldsymbol{\theta}_0;\ \boldsymbol{\theta}_1,\ s)\}]\end{aligned} \tag{3.3}$$

where $\mathbf{G} = \{G(\mathrm{d}\mathbf{x} \mid 1), \ldots, G(\mathrm{d}\mathbf{x} \mid m)\}$. Define $\boldsymbol{\theta}_0$ to be the value of $\boldsymbol{\theta}$, which maximizes the expected log likelihood (3.2) over all $\boldsymbol{\theta}$ satisfying $\mathrm{H}_0 : \boldsymbol{\beta}^{(1)} = \mathbf{0}_{p^{(1)}}$. A general definition of a stratified dependence measure based on information gain is

$$\begin{aligned}\varrho^2_{\mathrm{strat\,IG}} &= 1 - \prod_{s=1}^{m} \left[\frac{\exp\{-2\Phi(\boldsymbol{\theta}_1;\ \boldsymbol{\theta}_1,\ s)\}}{\exp\{-2\Phi(\boldsymbol{\theta}_0;\ \boldsymbol{\theta}_1,\ s)\}}\right]^{\omega_s}\\ &= 1 - \prod_{s=1}^{m} \left[\frac{\mathrm{M}_s(Y \mid \mathbf{X})}{\mathrm{M}_s(Y \mid \mathbf{X}^{(2)})}\right]^{\omega_s}\\ &= 1 - \frac{\mathrm{M}_{\mathrm{strat}}(Y \mid \mathbf{X})}{\mathrm{M}_{\mathrm{strat}}(Y \mid \mathbf{X}^{(2)})}\end{aligned} \tag{3.4}$$

where $\mathrm{M}_{\mathrm{strat}}(Y \mid \mathbf{X}^{(2)})$ and $\mathrm{M}_{\mathrm{strat}}(Y \mid \mathbf{X})$ denote the *stratified residual randomness* of Y under the hypotheses H_0 and H_1, respectively. They are the geometric means of the respective stratum-specific *residual randomnesses*.

When applying the construction principle for $\varrho^2_{\mathrm{strat\,IG}}$ to the stratified Cox model (1.1) we can make use of the fact that $h_{01}(t), \ldots, h_{0m}(t)$ are unspecified. That is, any set of strictly monotonic increasing functions $\phi_1(.), \ldots, \phi_m(.)$ can be applied to T in the corresponding strata and $\phi_s(T)$, $s \in \{1, \ldots, m\}$, will still give the same Cox regression coefficients as T. By accordingly choosing $\phi_1(.), \ldots, \phi_m(.)$ it can be ensured that the conditional distribution of $T^* = \phi_s(T)$ follows a Weibull distribution within each stratum $s \in \{1, \ldots, m\}$, so that $\mathrm{pr}\,(T^* > t \mid \mathbf{X} = \mathbf{x}, S = s) = \exp\{-t^{\alpha_s} \exp(\mu_s + \mathbf{x}\boldsymbol{\beta})\}$, for $\alpha_s > 0$. By definition, the stratified linear regression model (3.1) is valid for $\log(T^*)$, and the probability density function of the corresponding error term follows a standard extreme value distribution. Given $\mathbf{x}$ and s, the expected log likelihood for a standard extreme value distributed error term is

$$\Phi(\boldsymbol{\theta};\, \boldsymbol{\theta}_1,\, \mathbf{x},\, s) = \log(\alpha_s) + \frac{\alpha_s}{\alpha_{1s}}\, \gamma'(1) + b_s - \exp(b_s)\, \gamma\left(\frac{\alpha_s}{\alpha_{1s}} + 1\right),$$

where $b_s = \mu_s + \mathbf{x}\boldsymbol{\beta} - (\alpha_s/\alpha_{1s})\,(\mu_{1s} + \mathbf{x}\boldsymbol{\beta}_1)$. By accordingly making use of formulas (3.2), (3.3) and (3.4) we end up with the stratified dependence measure $\varrho^2_{\mathrm{strat\,W}} = 1 - \exp(-\Gamma)$, where the "W" in the subscript denotes the close relationship to the Weibull distribution.

$\varrho^2_{\mathrm{strat\,W}}$ can be estimated in an analogous manner as the measure for the unstratified case, ϱ^2_{W}. Since it can be shown that $\varrho^2_{\mathrm{strat\,W}}$ does not depend on the choice of μ_{1s} and $\alpha_{1s} > 0$, we set $\mu_{1s} = 0$ and $\alpha_{1s} = 1$, $s = 1, \ldots, m$. If the stratified censored survival data $\left\{(t_{si},\, c_{si},\, \mathbf{x}_{si});\ s = 1, \ldots, m,\ i = 1, \ldots, n_s, \sum n_s = n\right\}$ are given, then a stratified Cox regression model under H_1, that is no restrictions on $\boldsymbol{\beta}$, will yield the estimated vector $\tilde{\boldsymbol{\beta}}_{\mathrm{strat\,Cox},\,1}$ of regression coefficients so that $\tilde{\boldsymbol{\theta}}_1 = (\tilde{\boldsymbol{\beta}}^T_{\mathrm{strat\,Cox},\,1},\, \mathbf{0}^T_m,\, \mathbf{1}^T_m)^T$, where $\mathbf{0}_m$ and $\mathbf{1}_m$ are m-dimensional column vectors of zeros and ones, respectively. Estimating $\mathbf{G}$ and $\boldsymbol{\omega}$ by their corresponding empirical distributions yields an empirical expected log likelihood

$$\Phi(\boldsymbol{\theta};\, \tilde{\boldsymbol{\theta}}_1) = \sum_{s=1}^{m} \frac{n_s}{n} \left\{\frac{1}{n_s} \sum_{i=1}^{n_s} \Phi(\boldsymbol{\theta};\, \tilde{\boldsymbol{\theta}}_1,\, \mathbf{x}_{si},\, s)\right\} = \frac{1}{n} \sum_{s=1}^{m} \sum_{i=1}^{n_s} \Phi(\boldsymbol{\theta};\, \tilde{\boldsymbol{\theta}}_1,\, \mathbf{x}_{si},\, s)\,,$$

from where $\tilde{\boldsymbol{\theta}}_0 = (\mathbf{0}^T_{p^{(1)}},\, \tilde{\boldsymbol{\beta}}_0^{(2)T},\, \tilde{\mu}_{01}, \ldots, \tilde{\mu}_{0m},\, \tilde{\alpha}_{01}, \ldots, \tilde{\alpha}_{0m})^T$ has to be obtained numerically. If we plug in all these estimates into formula (3.3), we will get an estimate for Γ so that finally $\tilde{\varrho}^2_{\mathrm{strat\,W}} = 1 - \exp(-\tilde{\Gamma})$. Note that if the sampling scheme generates a strata frequency distribution different from the population of interest, then an alternative estimate for $\boldsymbol{\omega}$ could be used. By analogy the same is also valid for the distribution of $\mathbf{X}$ (KENT and O'QUIGLEY, 1988, p. 530).

If the numbers of observations per stratum, $n_1, \ldots, n_m$, are too small, then numerical convergence problems may occur when computing $\tilde{\varrho}^2_{\mathrm{strat\,W}}$. In our experience there should be at least five observations per stratum, although we would recommend $n_s \geq 10$, $s = 1, \ldots, m$. We found hardly any connection between the number of strata m and numerical convergence problems.

As in the unstratified case, an approximation $\varrho^2_{\mathrm{strat\,W,\,A}}$ for $\varrho^2_{\mathrm{strat\,W}}$ can be obtained by considering the error term of model (3.1) to follow a standard normal distribution. The information gain for the normal error model is given by $\Gamma = \omega_1 \log(\sigma^2_{01}) + \ldots + \omega_m \log(\sigma^2_{0m})$, so that $\varrho^2_{\mathrm{strat\,W,\,A}} = 1 - \prod_{s=1}^{m} (\sigma^2_{0s})^{-\omega_s}$. For $p^{(2)} = 0$ we find $\sigma^2_{0s} = 1 + \boldsymbol{\beta}_1^T \boldsymbol{\Omega}_s \boldsymbol{\beta}_1$, where $\boldsymbol{\Omega}_s$ is the covariance matrix of $\mathbf{X}$ in stratum s, $s = 1, \ldots, m$, and the estimate for $\varrho^2_{\mathrm{strat\,W,\,A}}$ is found by plugging in

$\tilde{\boldsymbol{\beta}}_{\text{strat Cox}, 1}$, $\tilde{\boldsymbol{\Omega}}_1, \ldots, \tilde{\boldsymbol{\Omega}}_m$, and $n_1/n, \ldots, n_m/n$ for $\boldsymbol{\beta}_1$, $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_m$, and $\omega_1, \ldots, \omega_m$, respectively. For $p^{(2)} > 0$ the expressions for $\sigma_{01}^2, \ldots, \sigma_{0m}^2$ are numerically cumbersome, so that there is no computational advantage in using the approximation $\varrho^2_{\text{strat W, A}}$ instead of $\varrho^2_{\text{strat W}}$ in this case.

An asymptotic $(1 - \eta)$-confidence interval for ϱ^2_{W} has been suggested by KENT and O'QUIGLEY (1988, p. 530), which after some appropriate modifications is also valid for the stratified dependence measures $\varrho^2_{\text{strat W}}$ and $\varrho^2_{\text{strat W, A}}$, respectively. It is defined as $[1 - \exp\{-\max(0, \tilde{\Gamma} - z_{1-0.5\eta}\sqrt{\tilde{v}/n})\};\ 1 - \exp\{-(\tilde{\Gamma} + z_{1-0.5\eta}\sqrt{\tilde{v}/n})\}]$, where $z_{1-0.5\eta}$ denotes the $(1 - 0.5\eta)$-quantile of the standard normal distribution, and

$$\tilde{v} = \tilde{\mathbf{c}}^T \tilde{\mathbf{V}}^{(11)} \tilde{\mathbf{c}} + \tilde{E}\,. \tag{3.5}$$

In order to define the single parts of formula (3.5) let $\ell(\boldsymbol{\beta}) = \ell_1(\boldsymbol{\beta}) + \ldots + \ell_m(\boldsymbol{\beta})$ denote the partial log likelihood of the stratified Cox regression model, let $\tilde{\mathbf{V}} = -\{n^{-1}\partial^2 \ell(\tilde{\boldsymbol{\beta}}_{\text{strat Cox}, 1})/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T\}^{-1}$ denote the inverse observed information matrix per observation, let $D_s = \tilde{\alpha}_{0s}/\tilde{\alpha}_{1s}$ and let $B_{si} = \tilde{\mu}_{0s} + \mathbf{x}_{si}^{(2)}\tilde{\boldsymbol{\beta}}_0^{(2)} - D_s(\tilde{\mu}_{1s} + \mathbf{x}_{si}\tilde{\boldsymbol{\beta}}_{\text{strat Cox}, 1})$. Now $\tilde{\mathbf{V}}^{(11)}$ is the submatrix of $\tilde{\mathbf{V}}$, which corresponds to the parameters $\boldsymbol{\beta}^{(1)}$ of interest. In the case of $\varrho^2_{\text{strat W}}$, $\tilde{\mathbf{c}} = \frac{2}{n}\sum_{s=1}^{m} D_s \sum_{i=1}^{n_s}\{1 - \gamma(D_s + 1)\exp(B_{si})\}\,\mathbf{x}_{si}^{(1)T}$, and $\tilde{E}$ is given by the sample variance of $\{-2\log(D_s) - 2D_s\gamma'(1) - 2B_{si} + 2\exp(B_{si})\ \gamma(D_s + 1);\ s = 1, \ldots, m,\ i = 1, \ldots, n_s\}$. On the other hand in the case of $\varrho^2_{\text{strat W, A}}$, $\tilde{\mathbf{c}} = -\frac{2}{n}\sum_{s=1}^{m} D_s \sum_{i=1}^{n_s} B_{si}\mathbf{x}_{si}^{(1)T}$, and $\tilde{E}$ is given by the sample variance of $\{-2\log(D_s) + D_s^2 + B_{si}^2;\ s = 1, \ldots, m,\ i = 1, \ldots, n_s\}$.

Based on asymptotic considerations of KENT (1983, p. 168), the expression $p^{(1)}\tilde{\Gamma}/\tilde{\Lambda}$ is suggested by KENT and O'QUIGLEY (1988, p. 531) as an approximate estimate of bias $(\tilde{\Gamma})$, where $\tilde{\Lambda}$ denotes the observed log partial likelihood ratio statistic. This correction of the inflation of the estimated information gain is directly applicable to the stratified case as well by replacing $\tilde{\Lambda}$ by the observed log stratified partial likelihood ratio statistic $\tilde{\Lambda}_{\text{strat}} = 2\{\ell(\tilde{\boldsymbol{\beta}}_{\text{strat Cox}, 1}) - \ell(\tilde{\boldsymbol{\beta}}_{\text{strat Cox}, 0})\}$. Hence, a bias-corrected estimate of the measure of dependence for the stratified Cox regression model is given by $\tilde{\varrho}^2_{\text{strat W, bc}} = 1 - \exp\{-\tilde{\Gamma}(1 - [p^{(1)}/\tilde{\Lambda}_{\text{strat}}])\}$. The corresponding approximation $\tilde{\varrho}^2_{\text{strat W, A, bc}}$ is defined analogously.

4. Simulation Study

There are situations where it is valid to account for confounder variables alternatively either by regression or by stratification. That is, both the unstratified partial R-squared measure ϱ^2_{W} and the stratified R-squared measure $\varrho^2_{\text{strat W}}$ are applicable and we expect their values to be similar. This was explored in the following simulation study.

Suppose that a survival time outcome T and three covariates, Z_1, Z_2 and Z_3, have been observed, and the proportional hazards assumption holds for all three covariates. Both Z_1 and Z_2 follow a standard normal distribution, whereas Z_3 is a qualitative factor with five levels. We want to measure the partial dependence between survival time T and the covariates Z_1 and Z_2 after accounting for the confounder Z_3 either by regression or by stratification, respectively. In the former case we will represent the 5-level categorical covariate Z_3 by four 0/1-coded dummy variables, D_1, D_2, D_3, and D_4, and after fitting the unstratified Cox proportional hazards regression model,

$$h(t \mid \mathbf{x}) = h_0(t) \exp\left(\mathbf{x}^{(1)}\boldsymbol{\beta}^{(1)} + \mathbf{x}^{(2)}\boldsymbol{\beta}^{(2)}\right), \tag{4.1}$$

with $\mathbf{X}^{(1)} = (Z_1, Z_2)$ and $\mathbf{X}^{(2)} = (D_1, D_2, D_3, D_4)$, we can compute the according estimate $\tilde{\varrho}^2_{\mathrm{W}}$. The other option is to stratify over the levels of Z_3, so that

$$h_s(t \mid \mathbf{x}) = h_{0s}(t) \exp\left(\mathbf{x}^{(1)}\boldsymbol{\beta}^{(1)} + \mathbf{x}^{(2)}\boldsymbol{\beta}^{(2)}\right), \qquad s = 1 \ldots 5\,, \tag{4.2}$$

with $\mathbf{X}^{(1)} = (Z_1, Z_2)$, $\mathbf{X}^{(2)}$ is empty and $S = Z_3$. After fitting the stratified Cox model (4.2) we can compute the according estimate $\tilde{\varrho}^2_{\mathrm{strat\,W}}$.

For the simulation study we chose the regression coefficients $\boldsymbol{\beta}^{(1)} = (1,\,1)^T$. In model (4.1) we used $\boldsymbol{\beta}^{(2)} = (\log(1.25), \log(1.5), \log(1.75), \log(2))^T$, which was accordingly translated into $h_{0s}(t) = [1 + 0.25(s-1)]\, h_0(t)$ for model (4.2), $s = 1, \ldots, 5$. For the sake of simplicity we set $h_0(t) = 1$ and assumed no censoring. A uniform distribution was assumed for the five levels of Z_3, thereby applying the constraint $n_s \geq 5$. Three different total sample sizes $n = n_1 + \ldots + n_5$ were considered, that is, 100, 200, and 500. The number of repetitions was set to 1000. The SAS (1996) software package was used for all numerical computations. The results can be found in Table 1.

The values for $\tilde{\varrho}^2_{\mathrm{W}}$ are slightly higher than that for $\tilde{\varrho}^2_{\mathrm{strat\,W}}$. The same is true when we compare their normal approximations $\tilde{\varrho}^2_{\mathrm{W,\,A}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W,\,A}}$, respectively. Note that for numerical reasons no value for the unstratified partial measure could be determined in 13 cases when $n = 100$. The fact, that for symmetric distributions $\tilde{\varrho}^2_{\mathrm{W,\,A}}$ is a slight overestimate of $\tilde{\varrho}^2_{\mathrm{W}}$, has been already noted by Kent and

Table 1

Mean results of 1000 repetitions. $\tilde{\varrho}^2_{\mathrm{W}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W}}$ measure the partial dependence between T and (Z_1, Z_2) after the effect of confounder Z_3 has been accounted for by regression and by stratification, respectively. $\tilde{\varrho}^2_{\mathrm{W,\,A}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W,\,A}}$ are approximations for $\tilde{\varrho}^2_{\mathrm{W}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W}}$, respectively. Sample size is denoted by n. The columns labelled "converged" show how often numerical convergence was successfully attained

n	$\tilde{\varrho}^2_{\mathrm{W}}$	$\tilde{\varrho}^2_{\mathrm{W,\,A}}$	converged	$\tilde{\varrho}^2_{\mathrm{strat\,W}}$	$\tilde{\varrho}^2_{\mathrm{strat\,W,\,A}}$	converged
100	0.65	0.67	987	0.63	0.65	1000
200	0.65	0.67	1000	0.64	0.66	1000
500	0.65	0.67	1000	0.64	0.66	1000

Table 2

Mean results of 1000 repetitions. $\tilde{\varrho}^2_{\mathrm{W}}$ measures the partial dependence between T and Z_1 after the effects of confounders (Z_2, Z_3) have been accounted for by regression. $\tilde{\varrho}^2_{\mathrm{strat\,W}}$ measures the partial dependence between T and Z_1, thereby accounting for the effects of Z_2 by regression and of Z_3 by stratification, respectively. $\tilde{\varrho}^2_{\mathrm{W,A}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W,A}}$ are approximations for $\tilde{\varrho}^2_{\mathrm{W}}$ and $\tilde{\varrho}^2_{\mathrm{strat\,W}}$, respectively. Sample size is denoted by n. The columns labelled "converged" show how often numerical convergence was successfully attained

n	$\tilde{\varrho}^2_{\mathrm{W}}$	$\tilde{\varrho}^2_{\mathrm{W,A}}$	converged	$\tilde{\varrho}^2_{\mathrm{strat\,W}}$	$\tilde{\varrho}^2_{\mathrm{strat\,W,A}}$	converged
100	0.48	0.50	984	0.46	0.47	993
200	0.48	0.50	999	0.47	0.48	1000
500	0.48	0.50	1000	0.48	0.49	1000

O'QUIGLEY (1988). It is not surprising that this is also valid for the stratified definition of the measure.

Next we considered Z_1 to be the only covariate of interest. That is, in the case of the unstratified model (4.1) $\mathbf{X}^{(1)} = (Z_1)$ and $\mathbf{X}^{(2)} = (Z_2, D_1, D_2, D_3, D_4)$, and in the case of the stratified model (4.2) $\mathbf{X}^{(1)} = (Z_1)$, $\mathbf{X}^{(2)} = (Z_2)$ and $S = Z_3$. The results of the corresponding simulations can be found in Table 2.

The structure of the results of Table 2 resembles that of Table 1. Again the unstratified results are slightly higher than the stratified ones. The occurrence of numerical non-convergence is more frequent in Table 2 than in Table 1. This is not surprising since from a numerical point of view the situation considered in Table 2 is much more delicate than that in Table 1. Interestingly, numerical troubles have been observed more often for the unstratified than for the stratified measures in both tables.

5. Example: Lung Cancer Data

The Veteran's Administration lung cancer data set (KALBFLEISCH and PRENTICE, 1980) consists of survival time and various covariates for 137 males with inoperable lung cancer. Only 9 of the 137 survival times are censored. In the following the continuous covariates age and Karnofsky performance index, and the qualitative four-level-covariate histology will be considered. The crossing of the estimated Kaplan-Meier survival curves formed by the four groups of histology (not shown) indicates an obvious violation of the proportional hazards assumption. We want to measure the dependence between survival of the veterans and the covariates age and Karnofsky performance index after accounting for the possible confounding effect of histology by stratification. The numbers of observations within the four strata are $n_1 = 35$, $n_2 = 48$, $n_3 = 27$, and $n_4 = 27$. After fitting of an according stratified Cox regression model we find $\tilde{\varrho}^2_{\mathrm{strat\,W}} = 0.309$. That is, the covariates age and Karnofsky performance index are able to explain a proportion of 0.309 of the randomness in the survival times, which is left after the effect of histology has been accounted for by stratification. The merely moderate accuracy of this statement is indicated by the rather wide 95% confidence interval, which ranges from 0.166 to 0.428. The bias is negligible, $\tilde{\varrho}^2_{\text{strat W, bc}} = 0.297$. The normal approximation $\tilde{\varrho}^2_{\text{strat W, A}} = 0.336$ is quite close.

If histology is ignored and an unstratified Cox model with covariates age and Karnofsky performance index is fitted to the data, then the corresponding unstratified dependence measure $\tilde{\varrho}^2_{\text{W}}$ will yield a value of 0.285. The difference to the result of the stratified analysis, $\tilde{\varrho}^2_{\text{strat W}} = 0.309$, is due to the fact that the assumptions for an unstratified analysis are incorrect. This difference is rather small in the current example, although it may become considerably large in cases with more severe violations of the proportional hazards assumption.

6. Discussion

The Cox model (Cox, 1972) has become the most popular regression model for analysing censored survival data in medical research. Guided by the ideal of the linear regression model with normal errors considerable efforts have been made to transfer the concepts for linear model diagnostics and linear model checking to the Cox model. Unfortunately, well-working and well-understood tools for the linear regression model may become quite hard to handle within the framework of the semiparametric Cox model. As shown in a detailed study by Schemper and Stare (1996), there is not a single, simple, easy to calculate, useful, easy to interpret R-squared measure for the Cox model. Quite on the contrary, various sensible definitions exist and each of them is restricted to a specific limited area of application.

The Kent and O'Quigley (1988) measure is certainly a reasonable choice in order to quantify overall or partial explained variation in the Cox model. Its area of application has been enlarged to the stratified Cox model in the present paper. The measure has a sound theoretical basis, and the numerical effort for its computation is in the range of the effort for estimating the Cox model itself. Another major motivation behind the Kent and O'Quigley (1988) measure is the invariance to the population quantity under independent censoring. This also holds in the stratified case. A drawback of the Kent and O'Quigley (1988) approach is its limited utilisability in the case of time-dependent covariates. Although the inclusion of such covariates into the Cox model is a straightforward and natural task, the corresponding adaptation of the measure will be only possible for the quite rare case of a population where the time-courses of these covariates are known over the entire time range, and even then reasonable generalizations of the expected log likelihoods (2.2) and (3.2) would be difficult, thereby saying nothing of the numerical computation of the corresponding estimates.

Finally, we would like to stress that situations in which the stratified Cox model could be used are quite common in practice, whereas the actual use of the stratified model is much less frequent. Part of the reason is probably the fact that proportional hazards assumption is often not thoroughly checked, and a common unstratified Cox model is wrongly applied. Other researchers again may recognise the need for dealing with non-proportional hazards but at the time avoid the usage of the Cox model at all, since they may find the results of the stratified model hard to interpret. We believe that having a measure that quantifies the partial effect of the covariates in the stratified model could help to remedy such practice. Non-proportionality of hazards is much more common than many researchers would want to recognise, and the stratified model provides a simple and efficient tool to deal with that (see for example O'QUIGLEY and STARE, 2002).

References

COX, D. R., 1972: Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.

HEINZL, H., 2000: Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model. *Computer Methods and Programs in Biomedicine* **63**, 71–76.

KALBFLEISCH, J. D. and PRENTICE, R. L., 1980: *The Statistical Analysis of Failure Time Data,* Wiley, New York.

KENT, J. T., 1983: Information gain and a general measure of correlation. *Biometrika* **70**, 163–173.

KENT, J. T. and O'QUIGLEY, J., 1988: Measures of dependence for censored survival data. *Biometrika* **75**, 525–534.

KULLBACK, S. and LEIBLER, R. A., 1951: On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

LAWLESS, J. F., 1982: *Statistical models and methods for lifetime data*, Wiley, New York.

LINFOOT, E. H., 1957: An informational measure of correlation. *Information and Control* **1**, 85–89.

O'QUIGLEY, J. and STARE, J., 2002: Proportional hazards models with frailties and random effects. To appear in: *Statistics in Medicine*.

REDDIEN, G. W., 1985: Newton-Raphson methods. In: S. Kotz and N. L. Johnson (eds.): *Encyclopedia of Statistical Sciences*, Volume 6. Wiley, New York, 210–212.

SAS, 1996: *The SAS System for Windows V6.12*, SAS Institute Inc., Cary, NC.

SCHEMPER, M. and STARE, J., 1996: Explained variation in survival analysis. *Statistics in Medicine* **15**, 1999–2012.

Received, August 2000
Revised, February 2002
Accepted, April 2002