On Choosing a Measure of Explained Variation for Cox Models

Janez Stare[1]

Abstract

There have been several attempts to extend the notion of explained variation to models for analyzing lifetime data, Cox models in particular. Most of the authors have concentrated on overcoming the difficulty of using the censored observations in the calculations, but did not pay enough attention to other, equally important issues. We present a list of properties that a measure of explained variation for the Cox model should have, and discuss these properties with reference to the measures by Kent and O'Quigley and by Korn and Simon. We show that only the approach by Kent and O'Quigley gives good results. We also stress the need for the use of such measures in survival studies.

1. Introduction

Since the aim of any regression model is to explain the variability of the dependent variable, it is desirable to generalize the definition of R^2, a customary measure of explained variation in linear regression, to other models. The most natural way of doing this is to apply the definition to such models. Unfortunately, this cannot be done easily. R^2 has several different interpretations of linear regression which lead to different statistics in more general models and there is no obvious way of deciding which definition to use. Some authors (Efron, 1978; Kvalseth, 1985; Nagelkerke, 1991) have tackled the problem of choosing the best definition and Kvalseth and Nagelkerke even produced a list of properties that such generalized R^2 should have. The field of survival analysis, which is the subject of this work, shows that obeying such 'general' properties may be difficult, if not impossible.

---

[1] Institute for Biomedical Informatics, Medical Faculty, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia.

Apart from the problem of choosing between different statistics, other difficulties occur when one is trying to define the concept of explained variation in survival analysis. For example, none of the definitions listed by Kvalseth can be directly applied to survival data. Furthermore, even some definitions that can be applied (like $R^2 = 1 - \exp(-LR/n)$, see Magee, 1990) have properties that are not desirable for such a measure. And the problems of meaningful interpretation and computational difficulty should also not be overlooked.

Taking all this into account, it is not surprising that the search for a measure of explained variation in survival analysis has only a brief history (Harrell et al., 1982; Harrell, 1986; Kent and O'Quigley, 1988; Korn and Simon, 1990; Schemper, 1990). It is interesting that different statistical packages (SPSS, NCSS, S-Plus) have implemented different measures, while some major packages (SAS, BMDP) have still not decided to include one. This is at least in part a consequence of the fact that no measure is widely agreed upon. Also, a search of literature using SCI has shown that these measures are only sporadically cited, which in view of the frequent use of survival analysis in medical literature confirms that this subject is still in development.

Criteria for choosing between measures are developed here. For illustration, two of the proposed measures are discussed, one by Kent and O'Quigley and the other by Korn and Simon. These and other measures have been empirically compared elsewhere (Stare, 1994; Schemper and Stare, 1996).

We start with a motivating example.

1.1 Survival after myocardial infarction

The study included patients who suffered myocardial infarction and were admitted to the Centre for Cardiovascular Diseases in Ljubljana, from May 1, 1982 to December 31, 1987. The study has been described in greater detail elsewhere (Stare, 1989; Kenda, Turk, and Stare, 1992). The aim of the study was to identify factors associated with the duration of survival. Data on 1040 patients were collected, but only 971 were complete and then used in proportional hazards regression analysis.

Results of the model fit are given in Output 2.1. All variables in the model are dichotomous, except for age, which is continuous. The following is a short explanation of the variables used in the table of estimates of maximum likelihood.

Heart arrythmia (ARRYTHM), took the value 1 if any kind of arrythmia was present, and 0 otherwise. *Cardiac failure* (CARD_FAI) was also a no-yes (0,1) variable. The variable *reinfarction* (REINFARCT) was 1 if the infarction in question was not the first infarction that a particular individual had suffered. *Age* entered the analysis as a continuous variable. All remaining variables were again dichotomous, having value 1 if a predictor was present, and 0 otherwise. Variables from ANT_TRAN to IMPOSSIB denote different localizations of the infarction; *anterior transmural, anterior nontransmural* and *impossible to locate*. Here, the inferior

infarction was the reference group. The next group of variables (NOREHAB to SPA) denotes different types of *rehabilitation*, with combined rehabilitation as the reference group. GLUCOSE stands for *disturbed tolerance for glucose*, and DIABETES is obvious. Here the group with neither diabetes or disturbed tolerance for glucose was the reference group.

Output 2.1. Analysis of Myocardial Infarction Data

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Per cent Censored
971	276	695	71.58

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	3673.059	3448.899	224.159
			with 14 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Pr > Chi-Square	Hazard Ratio
ARYTHM	1	0.409149	0.0015	1.50
INSUF	1	0.514078	0.0001	1.67
REINFRCT	1	0.359499	0.0267	1.43
AGE	1	0.029568	0.0001	1.03
ANT_TRAN	1	0.464313	0.0004	1.59
NT_NONT	1	0.485578	0.0482	1.62
IMPOSSIB	1	0.065668	0.8981	1.06
NOREHAB	1	0.958082	0.0706	2.60
NOTPRESC	1	0.916281	0.0810	2.50
AMBUL	1	-0.025794	0.9647	0.97
HOSPITAL	1	0.443442	0.4373	1.55
SPA	1	0.258553	0.6180	1.29
GLUCOSE	1	0.011207	0.9693	1.01
DIABETES	1	0.410995	0.0022	1.50

The study was concluded in 1992. Although more than 40% of patients were observed for more than 5 years, data is heavily censored, an indication of the low death-rate in the group.

As can be seen, the model is highly significant, but none of the variables has a very high hazard (HR) ratio, so the question of the usefulness of this model remained.

The model says that of two groups of people of equal age, with the first group having all the risk factors present while the second group has no risk factors, the first group has an almost 23-times-greater risk of dying than the second group (HR = $1.50 \cdot 1.67 \cdot 1.43 \cdot 1.62 \cdot 2.60 \cdot 1.50 = 22.6$). This becomes much higher if one compares older and younger groups.

A natural question to ask is: how good is this model? Is our prediction, based on the known values of prognostic variables, close to what really happens, or is there much left to explain?

The answer to such a question could be given by a measure analogous to the coefficient of determination in ordinary least squares regression. As mentioned, different measures of EV were proposed for the Cox model and Table 1 gives their values for the above example.

Table 1. Values of the measures for the infarction model.

	V1	V2	KS	Hc	Hcc	HL	KO
infarction	.15	.16	.20	.75	.50	.05	.45

V1 and V2 stand for Schemper's respective measures, KS for Korn and Simon's measure with squared error loss, Hc for Harrell's index of concordance, Hcc for the adjusted index of concordance, HL for Harrell's measure of explained likelihood, KO for Kent and O'Quigley's approximation.

It is obvious that something is wrong - the values differ too much to make any sense. We must therefore ask ourselves: what do we expect of such a measure? Since its crucial properties follow from the characteristics of the Cox model, we briefly discuss the model.

2. The proportional hazards model

The model was proposed by Cox (1972) and is often named after him. Before introducing the model, we must state some definitions.

Let T be a non-negative random variable representing the lifetimes[*] of individuals in a population. We will assume T to be continuous. In survival analysis we are interested in the function

$$S(t) = P(T \geq t),$$

where P denotes the probability of the event in parenthesis. This function is called the *survival function*. It is obvious that $S(t) = 1 - F(t)$, where $F(t)$ denotes the distribution function of the random variable T. In the study of survival time distributions, a useful concept is the *hazard function* defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function may be interpreted as the instantaneous probability that an individual will die in the interval $[t, t + \Delta t]$, provided that the event has not occurred before the beginning of this interval. It can be easily shown that $S(t)$ can be expressed in terms of $h(t)$ in the following way:

$$S(t) = \exp\left(-\int_0^t h(u)du \right). \tag{1}$$

Survival time cannot usually be directly related to explanatory variables via the regression model because of the presence of *censored data*. Data is said to be censored when exact lifetimes are known for only a portion of the individuals under study; and the remainder of the lifetimes are known only to exceed certain values[**]. Cox overcame this difficulty by using the hazard as the dependent variable in the regression model. His model takes the form

$$h(t|\mathbf{x}) = h_0(t)e^{\beta'\mathbf{x}}, \tag{2}$$

where β is the vector of the parameters, $\mathbf{x}$ is the vector of the independent variables and h_0 is the unspecified baseline hazard function. From (2) it follows that

---

[*] A *lifetime* does not necessarily represent the time until the death of an individual but can be the time up to the occurrence of some event that has elapsed from a well-defined starting point. Sometimes the terms *survival time* or *event time* are used.

[**] To be precise, such data is right censored. Other types of censoring exist, but we restrict our consideration to this type, which is the most common in practice.

$$\frac{h\left(t|\mathbf{x}_1\right)}{h\left(t|\mathbf{x}_2\right)} = e^{\beta'\left(\mathbf{x}_1 - \mathbf{x}_2\right)} ,$$

a property that gave the model its name. We note that this ratio is not time-dependent.

Assume we have n individuals in a sample and let $t_1, t_2, \ldots, t_k$ be ordered observed survival times, with $k \leq n$ since some observations may be censored. With $R(t)$ we shall denote the *risk set* at time t, the set of individuals that are still at risk just before t. The β coefficients of the model (2) are estimated by maximizing the so called *partial likelihood function*

$$PL = \prod_{i=1}^{k} \frac{\exp\left(\beta'\mathbf{x}_i\right)}{\displaystyle\sum_{j \in R(t_i)} \exp\left(\beta'\mathbf{x}_j\right)} ,$$

which depends only on β. It is because of this method of estimation of β that monotone transformations of time scale do not affect the estimates of β. For example, two models, illustrated in Figures 1 and 2, would have identical coefficients.

3. Properties of measures of EV for the Cox model

Given this property of the model, it is then natural to require the same of the measure. We can therefore state the first property which a measure of explained variation for the Cox model should have:

3.1 Monotone transformations of the time scale should have no effect on the measure

The second property comes from the fact that estimation of the coefficients in the Cox model is assymptotically independent of censoring.

3.2 The measure should be independent of censoring (consistency!)

While there is general agreement with respect to the second property (with different views about its actual importance), the first property is not viewed as necessary by all the authors. On the other hand, only one measure obeys the second property, and only one does not obey the first one. The second property, while desired, is difficult

to achieve, but the first one comes naturally from the consideration of the problem. The interested reader may consult original articles or Stare (1994) on this. As we have already said, we shall discuss only two measures here.

Figure 1. Two-group exponential model (HR=64)

Figure 2. Two-group Weibull model (HR=64)
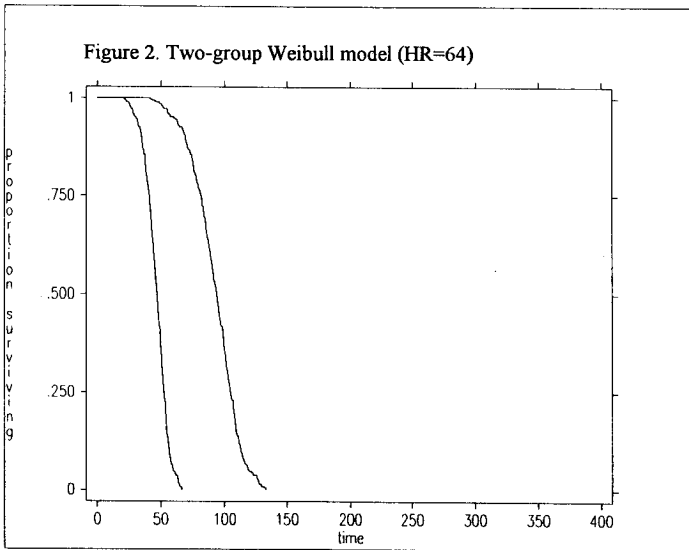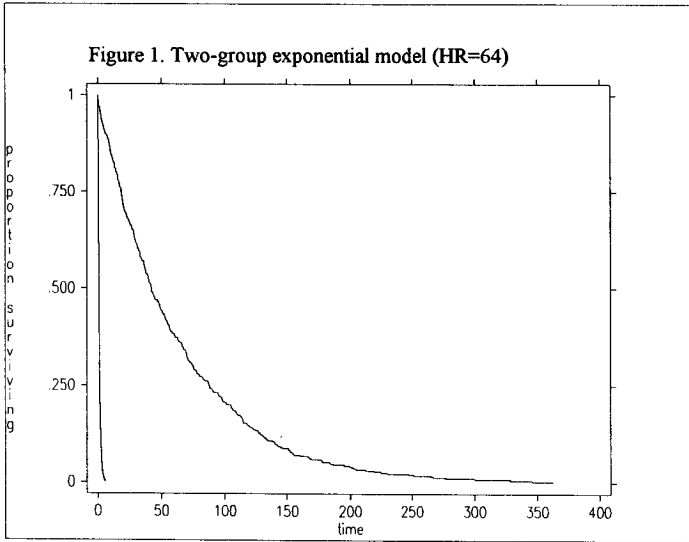
Other issues have been considered in Stare (1994) and in the original articles. These can be divided into two groups, those specific to the Cox model and those that are generally required. The specific characteristics of the measures for the Cox model are:

- the choice of the null model
- the effect of the sample size
- computational difficulty

As these are not considered very important we will omit their discussion here.

Of course there are some general properties that every measure of explained variation should possess. The most important seem to be:

- *known upper and lower limits* (1 and 0, say). Actually, known limits are only part of the problem. A measure can have such limits and still be too low or too high for all practical purposes (see Stare, 1994).
- *a measure must get higher as the model "fits better"*. This "obvious" requirement is not always obeyed.
- *a measure must have a clear interpretation*. This is of course a matter of opinion, but all things being equal, a measure with a clearer interpretation would be preferred.

Having stated the requirements we now proceed to discuss two measures, namely the measures by Kent and O'Quigley and by Korn and Simon. The reasons for choosing these two measures are the following:

- Kent and O'Quigley's measure is the only one that fulfils the above-stated requirements 1. and 2.
- Korn and Simon's measure does not fulfil any of the two requirements and therefore represents a different approach to solving the problem.

4. Kent and O'Quigley's measure of dependence

An information gain-based measure of dependence between the outcome variable and the explanatory variables was proposed by Kent and O'Quigley in their 1988 Biometrika paper.

They first note that the conditional distribution of T given a vector X of random variables $X_1, \ldots, X_p$ is specified only up to a monotone transformation of T. This is a consequence of using partial likelihood in estimating coefficients (see Kalbfleisch and Prentice, 1981). Therefore, for any monotone increasing function $\phi(\cdot)$, $T^* = \phi(T)$ also satisfies the Cox model with the same regression coefficients. One can then find

ϕ such that the base hazard function is of the form $h_0^*(t) = \alpha e^\mu t^{\alpha-1}$, for any μ and $\alpha > 0$. We shall denote by β a vector of coefficients in the Cox model and by $\mathbf{x}$ a vector of values taken by X. Since from (2)

$$h(t) = \alpha e^\mu t^{\alpha-1} e^{\beta x},$$

we have from (1)

$$S(t) = e^{-\int_0^t h(u)du} = e^{-\int_0^t \alpha e^\mu e^{\beta x} u^{\alpha-1} du} = e^{-t^\alpha e^{\mu+\beta x}}.$$

The variable $Y^* = \log T^*$ then follows a linear regression model

$$y^* = -\sigma\mu - \sigma\beta^{\mathrm{T}}\mathbf{x} + \sigma\varepsilon,$$

with $\sigma = \alpha^{-1}$, and the error ε follows the standardized extreme value distribution (see Lawless, 1982, pages 17-19 and 141). σ and μ are scale and location parameters, respectively.

Note that the above argument is theoretical and that finding a suitable transformation would in practice be impossible if no parametric form of baseline hazard was given. The point is that we can start to think in terms of the conditional distribution of Y given X being modelled by

$$Y = -\sigma\mu - \sigma\beta^{\mathrm{T}}X + \sigma\varepsilon,$$

where the error variate ε follows some specified distribution with a probability density function $f(\cdot)$. We shall restrict our consideration to the extreme value distribution, in which case $T = e^Y$ then follows the Weibull distribution.

Let $\Theta = (\beta, \mu, \sigma^2)$ denote the parameters of the model. Let $\Theta_1 = (\beta_1, \mu_1, \sigma_1^2)$ denote the true values of the parameters. We shall usually have $\beta_1 \neq 0$. Consider two hypotheses $H_0: \beta = 0$ and H_1: no restrictions on β. Denote the marginal distribution of X by $G(dx)$. Define Θ_0 to be the value of Θ maximizing the expected log likelihood

$$\Phi(\Theta, \Theta_1) = \iint \log\{f(y|x; \Theta)\} f(y|x; \Theta_1) dy G(dx) \tag{3}$$

over Θ satisfying H_0. Here $f(y|x; \Theta) = \alpha f(\alpha y + \mu + \beta^{\mathrm{T}}x)$ with $\alpha = \sigma^{-1}$. Kullback and Leibler (1951) defined the *information gain* by

$$\Gamma = 2\big(\Phi(\Theta_1; \Theta_1) - \Phi(\Theta_0; \Theta_1)\big). \tag{4}$$

As a measure of dependence between Y and X, Kent (1983) proposed

$$\rho_{IG}^2 = 1 - e^{-\Gamma}.$$

This is motivated by the fact that in the linear regression model with normal errors $\left(f(y) = (2\pi)^{-1/2} \exp\left(- y^2/2\right)\right)$, ρ_{IG}^2 reduces to the squared multiple correlation coefficient between Y and X.

For the Weibull model, the expected log likelihood can be written in closed form. In general, estimation of information gain is quite involved. Kent and O'Quigley proposed an approximation for a Weibull model to be

$$\rho_{W,A}^2 = \frac{A}{A+1}, \tag{5}$$

where $A = \beta^T \Omega \beta$ and Ω is the covariance matrix of X. This is estimated by the sample covariance matrix S.

Kent and O'Quigley note the following properties of their measure:

i) $0 \leq \rho_{IG}^2 < 1$, $\rho_{IG}^2 \to 1$ when $\|\beta\| \to \infty$. ρ_{IG}^2 will be 0 if H_0 is true, but this may not hold for a misspecified model.

ii) ρ_{IG}^2 is invariant under linear transformations of Y and X. It depends only on the regression coefficients β and on the marginal distribution of X, but not on μ or σ.

iii) For a sample of n independent observations $\left\{ (x_i, y_i); i = 1, \ldots, n \right\}$ without censoring, the estimate of Γ is just n^{-1} times the usual log likelihood ratio statistic for testing H_0 versus H_1. Then ρ_{IG}^2 becomes

$$\rho_{IG}^2 = 1 - e^{(-LR/n)}.$$

The same measure has been defined by Maddala (1983) and discussed by Magee (1990) for the binomial logit model and by Schemper (1992) for the Cox model. He showed that the measure depends on censoring and on the validity of the model's assumptions.

It is of course obvious that (5) satisfies both requirements stated on page 7, but the same can seen for the information gain measure.

It is clear from (3) and (4) that the value of Γ will depend on the values of

$$\log\left\{ f(y|x; \Theta_1) / f(y|x; \Theta_0) \right\} \tag{6}$$

These will be greater with greater difference between the two probability spaces $(\Omega, \mathrm{F}, \mu_0)$ and $(\Omega, \mathrm{F}, \mu_1)$, where the probability measures are represented by the respective densities $f(y|x; \Theta_i)$. Or, in the words of Kullback and Leibler (1951), if $H_i, i = 0, 1$, is the hypothesis that x was selected from the population whose probability measure is μ_i, then (6) is the information in x for discrimination between H_0 and H_1.

5. Korn and Simon's measure

We give here an outline of the 'loss function' approach by Korn and Simon (1990).

If we predict a survival time of an individual to be p and his actual survival time is t, then some loss is incurred. Such loss is a function of p and t which we denote as $L(t, p)$ and call a *loss function*. Many different forms of loss function can be used, the only requirement being that with a greater difference between t and p, the loss is also greater. Korn and Simon consider several such loss functions, of which we will discuss the following one:

$$L(t, p) = (t^* - p)^2, \tag{7}$$

where $t^* = \min(t, T_0)$ and T_0 is the censoring time above which we have no interest in the survival process. A prediction $p = T_0$ should therefore be interpreted as a prediction of 'survival greater than or equal to T_0'. With an observed time $t \geq T_0$, such a prediction incurs no loss. With $T_0 = \infty$ (7) becomes

$$L(t, p) = (t - p)^2. \tag{8}$$

For a given loss function, *expected risk* is defined for a survival curve S by

$$R_L(S(\cdot)) = \min_p \int L(t, p) dF(t),$$

where $F(t) = 1 - S(t)$ is the distribution function of the random variable T^2. Korn and Simon call a p achieving this minimum the 'optimal predictor'. The optimal predictors for the two loss functions (7) and (8) are $E(T^*)$ and $E(T)$.

---

[2]We give here an intuitive explanation of the above definition:

Let p be a predicted survival time for a group of n subjects which are the same as far as the values of covariates are concerned. Let t_i be the actual survival time of subject i. Then the loss incurred from predicting p instead of t_i is $L(t_i, p)$. We are interested in the sum (integral) of all

The problem with survival analysis data is that losses usually cannot be calculated, because of the censored observations. To calculate the expected risk for a subject with given covariate vector x_i, Korn and Simon use the predicted survival curve $S(t|x_i)$, which is calculated from what is assumed to be a correctly specified model. They define the explained variation of the survival model to be the proportional reduction in risk obtained by using the model for prediction over the null model:

$$\text{explained variation} = \frac{R_L[S_0(\cdot)] - \frac{1}{n}\sum_{i=1}^{n} R_L[S(\cdot|x_i)]}{R_L[S_0(\cdot)]}.$$

They define the null model to be

$$S_0(t) = \frac{1}{n}\sum_{i=1}^{n} S(t|x_i).$$

which is a mixture distribution of the modelled survival distributions.

The following properties of this measure can be seen from the definition:

- first, censored observations represent no problem in calculating this measure.
- second, using the loss function (7), a time range of interest can be specified.
- third, for different values of T_0, the measure will also be different.
- fourth, any transformation of time scale will affect the measure.
- fifth, the measure is clearly model dependent.

The fourth property clearly contradicts requirement 1. of section 3. Requirement 2. will be satisfied only if the expected risks can be calculated for $T_0 = \infty$. In practice, data is often censored because the data is analyzed before all deaths in the sample occur, so the upper limit of T_0 is imposed by the data at hand, making the measure dependent on censoring.

---

losses $\sum L(t_i, p)$ and we want to make this sum as small as possible. We would regard a value of p which minimizes this sum as the best possible predictor for a given group of subjects. The above sum of losses of course depends on the distribution of t_i. If t_i are close to p it is small; if far from p. then it is large. So, in the definition (1.9) the integral sign represents the sum and the distribution of T is represented by $dF(t)$.

6. Conclusions

The most important conclusion is: use only the Kent and O'Quigley measure.

Second, the quest for a measure of explained variation in survival analysis is not at an end. An important feature of the Cox model is the possibility of including time-dependent covariates and these cannot be used with any of the measures mentioned. The measure introduced recently by O'Quigley and Flandre (1994) promises to solve this problem while preserving both of the two most important properties.

On the other hand, people may not find these approaches too appealing because of the difficulties with their interpretation. In this respect, Schemper's recent proposal (see Schemper and Stare, 1996) is certainly a candidate to fare better.

One final word about the necessity of regularly using measures of predictive power in survival studies. It is of course well-known that statistical significance has nothing to do with the predictive power of prognostic variables. Still, too many models are built relying solely on statistical significance. More importantly, explained variation is sometimes confused with the goodness of fit (see a good discussion on this in Korn and Simon, 1990a and 1990b). We note that correct specification of the model only ensures that predictions will be correct on average, but does not imply precise individual predictions.

References

[1]   Cox D.R. (1972): Regression models and life-tables. *J Roy Statist Soc B*, **34**, 187–220.

[2]   Efron B. (1978): Regression and ANOVA with zero-one data: Measures of residual variation. *JASA*, **73**, 113–21.

[3]   Harrell F.E. (1986): The PHGLM procedure. In SUGI Supplemental Library User's Guide, Version 5 ed., Ed. Hastings, R.P., Cary, North Carolina: SAS Institute Inc., 437-466.

[4]   Harrell F.E., Califf R.M., Pryor D.B., Lee K.L., and Rosati R.A. (1982): Evaluating the yield of medical tests. *JAMA*, **247**, 2543–2546.

[5]   Kalbfleisch J.D. and Prentice R.L. (1981): *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

[6]   Kenda M.F., Turk J., and Stare J. (1992): Death after acute myocardial infarction - nine years of different aspects. *New Trends Arrhyt*, **8**, 375–381.

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

[8]   Kent J.T. and O'Quigley J. (1988): Measures of dependence for censored survival data. *Biometrika* , **75**, 525–34.

[9] Korn E.L. and Simon R. (1990a): Explained residual variation, explained risk, and goodness of fit. *The American Statistician*, **44**, 250–253.

[10] Korn E.L. and Simon R. (1990b): Measures of explained variation for survival data. *Statistics in Medicine*, **9**, 487–503.

[11] Kullback S. and Leibler R.A. (1951): On information and sufficiency. *Ann Math Stat*, **22**, 72–86.

[12] Kvalseth T.O. (1985): Cautionary note about R2. *The American Statistician*, **39**, 279–85.

[13] Lawless J.F. (1982): *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

[14] Maddala G.S. (1983): *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.

[15] Magee L. (1990): R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, **44**, 250–253.

[16] Nagelkerke N.J.D. (1991): A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–2.

[17] O'Quigley J. and Flandre P. (1994): Predictive capability of proportional hazards regression. *Proc. Natl. Acad. Sci. USA*, **91**, 2310–2314.

[18] Schemper M. (1990): The explained variation in proportional hazards regression. *Biometrika*, 77, 216–8.

[19] Schemper M. (1992): Further results on the explained variation in proportional hazards regression. *Biometrika*, **79**, 202–204.

[20] Schemper M. and Stare J. (1996): Explained variation in survival analysis. (To appear in *Stat Med*).

[21] Stare J. (1989): Statisticna analiza dobe prezivetja in njena uporaba pri bolnikih po prebolelem miokardnem infarktu. Ljubljana: Medicinska fakulteta, Master degree thesis (in Slovene).

[22] Stare J. (1994): Measures of explained variation in survival analysis. Doctoral Dissertation. Medical Faculty, Ljubljana.