Goodness of fit criteria for survival data

Martina Müller[1]

Institute for Medical Statistics and Epidemiology, IMSE
Technical University Munich, Ismaningerstr.22, 81675 München, Germany

ABSTRACT

The definition of an appropriate measure for goodness-of-fit in case of survival data comparable to R^2 in linear regression is difficult due to censored observations. In this paper, a variety of answers based on different residuals and variance of survival curves are presented together with a newly introduced criterion. In univariate simulation studies, the presented criteria are examined with respect to their dependence on the value of the coefficient associated with the covariate; underlying covariate distribution and censoring percentage in the data. Investigation of the relations between the values of the different criteria indicates strong dependencies, although the absolute values show high discrepancies and the criteria building processes differ substantially.

[1] muellerin2001@web.de

1 Introduction

A major interest of survival analysis is the investigation and rating of prognostic factors for specific diseases. Survival analysis as time-to-event models is often realised by semiparametric Cox regression which does not allow for direct computation of a measure of goodness-of-fit such as R^2 for linear regression due to incomplete observation times i.e. censored failure times. Several attempts were made to establish an at least comparable measure. An appropriate measure should represent the difference between the real data and the predicted values of the model and be dependent on the estimated coefficients. In addition, it should be able to be interpreted as a percentage of variation in the data that explained by the model. Some of the proposed measures have recently been corrected to reduce dependence on the percentage of censoring in the data.
The aim of this paper is to investigate the latest measures along with a newly introduced variant in univariate simulation studies. They will be analysed with respect to their dependence on underlying covariate distribution, strength of the covariate's influence, which is the associated coefficient, and censoring percentage in the data. The absolute values of the different measures show high discrepancies. As they all are constructed for the same purpose, the associations between the values of the different measures resulting from simulated data were examined. It was found that they are strongly related to each other although they are based on different outcomes of survival analysis.
In section two, the background of survival analysis, a general definition of R^2 and desirable properties of an appropriate measure are outlined. In section three, the definitions of existing criteria along with a newly introduced variant, which measure the goodness-of-fit in survival analysis, are presented. Simulation results are shown and discussed in section four, and in section five, an application to real data is given.

2 Background

2.1 Survival analysis

The main interest of survival analysis is usually the probability to survive until a chosen point on the time axis. This is described by the survivor function $S(t)$. The cumulated probability to die until time t, $F(t)$, is related to the survivor function by:

$$S(t) = 1 - F(t)$$

The hazard function $\lambda(t)$ represents the instantaneous probability to die at time $t+\delta t$ for a subject that survived at least until t. The cumulated hazard $\Lambda(t) = \int_0^t \lambda(s)ds$ is related to the survival function by:

$$S(t) = \exp(-\Lambda(t))$$

The estimation of the survival function can be realised nonparametrically by Kaplan-Meier estimator (Kaplan & Meier, 1958). For each point of the time axis, the probability of death is calculated as the number of events d_i relative to the number $R(t_i)$ of subjects at risk at time t_i. Hence, censored failure times enter the estimation as a reduction in the corresponding number of subjects at risk. The Kaplan-Meier estimator is written:

$$S_{KM}(t) = \prod_{i=0}^{t}(1 - \frac{d_i}{R(t_i)})$$

A Kaplan-Meier survival curve is therefore a step function over time with steps at each time an event occurs, i.e. each failure time. In case of discrete covariates, survival curves for different factor levels can be calculated and compared by logrank test which gives an indication of the relevance of the factor, i.e. whether survival in the groups significantly differs.
Continuous as well as discrete covariates can be handled by the semiparametric Cox proportional hazards model (Cox, 1972). This model assumes a general baseline hazard $\lambda_0(t)$ for all subjects, given that all covariates have outcome zero. This baseline hazard is arbitrary over time t. Nonzero covariate values result in a constant shift of this baseline hazard over time.
Some software packages, such as S-plus, use mean covariate values for the calculation of the baseline hazard instead of zero which results in a constant shift of this function. The knowledge of the used reference is only important for the interpretation of the resulting baseline hazard function.
The assumption of a constant shift of the baseline hazard by covariate values is called *proportional hazards* and must be checked before interpreting the results of a Cox regression. If it is not fulfilled, a different model must be applied, e.g. one allowing for time-varying coefficients. For a valid Cox model, the formula for the hazard function is given as:

$$\lambda(t|x) = \lambda_0(t)\exp(\beta' X)$$

Estimation in Cox regression is based on the *partial likelihood function* which is the first part of the full likelihood and independent of the underlying baseline hazard (Cox, 1975). It has been shown to have similar features as the full likelihood although some information is lost by reducing the full likelihood to its first term. The loss of information is not negligible for small data sets and for informative censoring. It is usually assumed that censored observations do not contribute additional information to the estimation. This is the case, if censoring is independent of the survival process. Otherwise, censoring is informative and estimation via partial likelihood is biased.
If the data set is large enough and censoring is uninformative, estimation in proportional hazards regression is established by maximising the logarithmised partial likelihood, which is:

$$\ln PL(\beta, X) = \sum_{Y_i uncensored} \left(X_i\beta - \ln \sum_{t_j \geq t_i} \exp(\beta' X_j) \right)$$

For tied failure times, a correction must be introduced. Breslow (1974) proposed the following correction of the partial likelihood function:

$$\ln PL(\beta, X) = \sum_{Y_i uncensored} \left(s_i\beta - d_i \left[\ln \sum_{t_j \geq t_i} \exp(\beta' X_j) \right) \right]$$

Herein, s_i is the sum of the covariates of all individuals and d_i is the total number of individuals failing at the ith failure time.
Maximisation is realised by setting the score function, which is the first derivative of the logarithmised partial likelihood, to zero. The score function is:

$$U(\beta, X) = \sum_{Y_i uncensored} \left(X_i - \frac{\sum_{t_j \geq t_i} X_j \exp(X_j\beta)}{\sum_{t_j \geq t_i} \exp(X_j\beta)} \right)$$

Cox regression has become the standard for survival analysis. However, in practise, the assumption of proportional hazards is rarely checked although a wide choice of models allowing for non-proportional hazards by accounting for time-dependent effects $\beta(t)$ has been proposed. An example is given by Berger et al. (2003) who model time-dependent effects by fractional polynomials.

2.2 Goodness-of-fit criteria

A range of measures of goodness-of-fit and their application for different settings are described by Kvalseth (1985). In ordinary linear regression, R^2, based on the residual sum of squares, is often the chosen measure for judging the fit of a model. It results from the decomposition of sums of squares where SST is defined as total sum of squares, SSR as the residual sum of squares and SSM as the sum of squares explained by the model:

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}$$

However, this measure cannot be used for proportional hazards regression as the outcome contains incomplete failure times. The definition of residuals for these models is more complicated and a variety of answers is available. Some of which have been used to create criteria measuring the influence of the covariates. The resulting criteria definitions are presented, along with others, which are not based on residuals, and a newly introduced variant, in the next chapter.
General desirable properties of a measure similar to R^2 in linear regression have been formulated by Kendall (1974). These are:

- $R^2 = 0$ in absence of association
- $R^2 = 1$ for perfect predictability
- R^2 should increase with the strength of association

These three stipulations should be checked within the simulation section for the presented criteria. Other desirable properties are that the value should increase with the absolute value of the coefficient associated with the examined covariate, and that the measure should not be influenced by the percentage of censoring in the data. The latter of these properties is not easy to solve. Although some of the existing measures have been corrected recently, the simulation results for the new measures indicate that dependencies on the censoring percentage are weaker but still exist.

3 Criteria definitions

3.1 Measures based on martingale and deviance residuals

Martingale residuals are defined as the difference between the cumulative hazard assigned to an individual with failure time t_i and its observed status, $\delta_i = 0$ censored, $\delta_i = 1$ event. Martingale residuals are written as follows (Therneau, Grambsch, Fleming, 1990):

$$M_i = \delta_i - \Lambda(t_i)$$

As the cumulative hazard $\Lambda(t_i)$ has no upper limit, these residuals range between 1 and $-\infty$. However, the sum of all martingale residuals is always 0. $\Lambda(t_i)$ is the number of expected events per individual failing at t_i according to the model. In a perfect model, which is defined as a perfect prediction for all individuals, all martingale residuals are 0 and uncorrelated to each other. There is a slight negative correlation in all other models due to the property that they sum up to 0.
A high $\Lambda(t_i)$ can be interpreted as a high indication of death and will result in a highly negative martingale residual. According to the model, these individuals were under observation for too long. As $\Lambda(t)$ increases with time, the residuals will tend towards increasingly negative values for longer observation times and have a highly skewed distribution.
Normalised transformations of martingale residuals have therefore been defined. These transformed versions are called *deviance residuals*:

$$devM_i = sgn(M_i)\sqrt{-2(M_i + \delta_i \ln(\delta_i - M_i))}$$

This definition resembles the definition of deviance residuals in Poisson regression, but as the nuisance parameter, the unspecified baseline hazard $\lambda_0(t)$ is still involved, and the squared residuals do not exactly sum up to

the deviance of the model, as they would do for Poisson regression (Venables & Ripley, 1997). Hence, the sum of squared residuals cannot be interpreted as the total deviance of the model.
Martingale residuals can be used to detect the functional form of a covariate. Outlier screening can be performed by plotting either kind of residual against time but this may be established more easily by using deviance residuals.
Especially for covariate models with high coefficients and low censoring percentage, martingale residuals tend towards extreme negative values.
A common property of these residuals is their high dependence on the amount of censoring in the data. The survival function is always calculated with respect to the amount of subjects at risk. Censoring by means of the end of a study will always increase the probability that a subject who is expected to have a long expected survival time is censored in comparison to a subject who is expected to have a short expected survival time. There are therefore usually more censored observations at the end of the study. The survival function only has steps at times that are marked by failures. If the last point on the time axis is not a failure, the survival function flattens earlier than it would do if the last observation is a failure. This is because only the number of individuals at risk at the last observed failure is taken into account. Flattening survival also results in flattening cumulative hazard and therefore less extreme residuals. On the other hand, extreme residuals occur for high cumulative hazards. Consequently, these can be obtained in cases of low censoring or high values of the product of covariates and associated coefficients, i.e. the linear predictor.
Adopting the idea of a goodness-of-fit criterion based on residual sums of squares, comparable to R^2 in linear regression, will often result in the preference of a null model over any covariate model. This is because the squared extreme residuals resulting from high values of the linear predictor increase the sum of squares to such extremes that R^2, which includes the difference between the sum of squares of residuals from null and covariate model, can even yield high negative values. The assumption of having normally distributed residuals, which is needed for the application of R^2, is simply not fulfilled because the distribution of martingale residuals is more exponentially shaped. This also applies to the more normally distributed deviance residuals - weaker but still visible.
Hence, a different definition for an appropriate criterion is needed. Stark (1997) proposed measuring the mean absolute differences between residuals of null and the covariate model and setting the sum relative to the mean of absolute residuals of the null model. Thus, with $M_{i|x}$ as martingale residual in the covariate model and M_i as residual of the null model, the new measure is written:

$$K_{m.norm} = \frac{\frac{1}{n}\sum_i \left|M_i - M_{i|x}\right|}{\frac{1}{n}\sum_i |M_i|} = \frac{\sum_i \left|\Lambda_i - \Lambda_{i|x}\right|}{\sum_i |M_i|}$$

Note that the observed status cancels in the counter of the fraction when calculating absolute differences.
For deviance residuals an analogous measure can be built:

$$K_{d.norm} = \frac{\frac{1}{n}\sum_i \left|devM_i - devM_{i|x}\right|}{\frac{1}{n}\sum_i |devM_i|} = \frac{\sum_i \left|devM_i - devM_{i|x}\right|}{\sum_i |devM_i|}$$

This new definition does not completely circumvent problems arising from these types of residuals. Although negative values are avoided, the differences between the residuals of the two models may still get very high, i.e. the difference between two residuals resulting from the two applied models may be larger than the residual of the null model. In this case, often the measures can exceed the maximum value allowed for a goodness-of-fit criterion, which is 1 (Kendall, 1975). Simulation studies presented later showed that these problems arise especially for data with low censoring percentage and high discrepancies of the linear predictor. The latter case occurs if the true underlying covariate distribution allows for high variance of covariate values in combination with a high coefficient.
Correction is difficult, but will be the object of further research.

3.2 Measures of variation in survival

As survival can be taken as a major point of interest, criteria have been proposed that are based on the survival function. Initially, the absolute distance or the mean squared distance between the survival curves of a null model obtained through Kaplan-Meier estimation and a covariate including Cox model were measured (Schemper, 1990).
These were later improved by measuring the weighted reduction of variance in the survival processes (Schemper & Henderson, 2000). The variance of the individual survival process at time t is defined as $S(t)\{1 - S(t)\}$ for a null model and $S(t|X)\{1 - S(t|X)\}$ for the covariate model (Schemper & Henderson, 2000). The mean absolute deviation measures are defined as $2S(t)\{1 - S(t)\}$ and $2S(t|X)\{1 - S(t|X)\}$. Measures of predictive accuracy integrated over the full follow-up range are weighted by a function of time to reduce dependence on censoring and the factor 2 is dropped as it cancels in the resulting criteria definitions. Hence:

$$D(\tau) = \frac{\int_0^\tau S(t)\{1 - S(t)\} f(t)dt}{\int_0^\tau f(t)dt}$$

$$D_x(\tau) = \frac{\int_0^\tau E_X\left[S(t|X)\{1 - S(t|X)\}\right] f(t)dt}{\int_0^\tau f(t)dt}$$

The measure V of the relative gain is then formulated. It is written as the difference between the variance in survival for the null model, D, and its expectation for the covariate model, D_x, relative to that of the null model:

$$V(\tau) = \frac{D(\tau) - D_x(\tau)}{D(\tau)}$$

An alternative formulation is defined using weighted relative gains. The weighting functions are moved as follows:

$$V_W(\tau) = \frac{\int_0^\tau \frac{S(t)\{1-S(t)\}-E_X[S(t|X)\{1-S(t|X)\}\}}{S(t)\{1-S(t)\}} f(t)dt}{\int_0^\tau f(t)dt}$$

For estimation, $S(t)\{1 - S(t)\}$ and $E_X[S(t|X)\{1 - S(t|X)\}]$ must be split into three terms at each distinct death time $t_{(j)}$ as there are individuals still alive (line 1 below), individuals that died before $t_{(j)}$ (line 2) and those who are censored before $t_{(j)}$ (line 3). The mean absolute distance measure D is therefore estimated by $\hat{M}(t_{(j)})$ as follows:

$$\begin{aligned}\hat{M}(t_{(j)}) = \frac{1}{n}\sum_{i=1}^{n} \Big[& I(t_i > t_{(j)})\left\{1 - \hat{S}(t_{(j)})\right\} \\ & + \delta_i I(t_i \le t_{(j)})\hat{S}(t_{(j)}) \\ & + (1-\delta_i)I(t_i \le t_{(j)})\left\{(1 - \hat{S}(t_{(j)}))\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)} + \hat{S}(t_{(j)})\left(1 - \frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)}\right)\right\}\Big]\end{aligned}$$

D_x is calculated the same way as D only with survival estimates $\hat{S}(t_{(j)})$ replaced by $\hat{S}(t_{(j)}|X)$, which are obtained from a Cox model.
In the next step, the weights are calculated from the potential follow-up distribution, also called *reverse Kaplan-Meier*, which is estimated like a Kaplan-Meier estimator for survival, but with the meaning of the status indicator δ reversed (Schemper & Smith, 1996 and Altman et al., 1995). The reverse Kaplan-Meier function is denoted $\hat{G}$. With d_j being the number of deaths at time $t_{(j)}$, the weights at time $t_{(j)}$ are defined as follows:

$$w_j = \frac{d_j}{\hat{G}(t_{(j)})}$$

Hence, the estimate $\hat{V} = (\hat{D} - \hat{D}_x)/\hat{D}$ is calculated with:

$$\hat{D} = \frac{\sum_j w_j \hat{M}(t_{(j)})}{\sum_j w_j} \qquad \text{and} \qquad \hat{D}_x = \frac{\sum_j w_j \hat{M}(t_{(j)}|x)}{\sum_j w_j}$$

And V_W is:

$$\hat{V}_W = \frac{\sum_{j=1}^{m} w_j \frac{\hat{M}(t_{(j)}) - \hat{M}(t_{(j)}|x)}{\hat{M}(t_{(j)})}}{\sum_j w_j}$$

Simulation studies showed that the value of V_W is always slightly smaller than V.

3.3 Measures based on Schoenfeld residuals

A different method for judging a model is based on Schoenfeld residuals (Schoenfeld, 1982). These measure the model's accuracy in a different way. At all complete failure times, the true covariates assigned to an individual failing are compared to the expected value of the covariate under the model, assuming the model is true. Hence, the idea is completely different from the measures presented so far. While martingale residuals and the variance of the survival curves are calculated conditioning on the covariates, these residuals investigate covariate values conditioning on time. The expected values of the covariates are calculated with respect to the probabilities assigned to the values by the model. With $r_i(t)$ as indicator whether individual i is still at risk at time t, the Schoenfeld residuals are defined as follows:

$$\begin{aligned} r_{sch_i}(\beta) &= X_i(t_i) - E(X_i(t_i)|\beta) \\ &= X_i(t_i) - \sum_{j=1}^{n} X_j(t_i)\pi_j(\beta, t_i) \\ &= X_i(t_i) - \sum_{j=1}^{n} X_j(t_i) \frac{r_j(t_i)\exp(\beta' X_j(t_i))}{\sum_{j=1}^{n} r_j(t_i)\exp(\beta' X_j(t_i))} \end{aligned}$$

Consequently, when calculating these residuals, the result will be a matrix with the number of rows equalling the number of events and a column per covariate.

In order to define a measure for the goodness-of-fit of the model a modified version of residuals is needed for a null model. This is difficult because the residual is based on the covariates. For the null model, the covariate is supposed to have no influence and can be assigned to the individuals arbitrarily. Therefore, residuals for the null model are obtained by replacing the probability $\pi_j(\beta, t_i)$ in the upper definition for the covariate model by $\pi_j(0, t_i) = r_j(t_i) / \sum_{j=1}^{n} r_j(t_i)$. In this way, all covariate values have the same probability and there is no preference for values as in a covariate model. A first measure for the goodness-of-fit has been formulated based on squared residuals (O'Quigley & Flandre, 1994):

$$R^2_{OF} = \frac{\sum_{i=1}^{n} r^2_{sch_i}(0) - \sum_{i=1}^{n} r^2_{sch_i}(\beta)}{\sum_{i=1}^{n} r^2_{sch_i}(0)}$$

$$= 1 - \frac{\sum_{i=1}^{n} r_{sch_i}^2(\beta)}{\sum_{i=1}^{n} r_{sch_i}^2(0)}$$

The next problem is that the residuals are calculated per covariate and it is difficult to judge multivariate models. The idea of the *prognostic index* (Andersen et al., 1983) has therefore been adopted as each patient's outcome is dependent of the combination of all covariates. The prognostic index is defined as:

$$\eta(t) = \beta' X(t)$$

Hence, the definition of a criterion, which can also be applied to a multivariate model, is based on Schoenfeld residuals multiplied by the vector of covariates.
To reduce dependencies on the censoring percentage in the data, the squared residuals are weighted by the height of the increment of the marginal survival curve at the corresponding point on the time axis. This is obtained from the marginal Kaplan-Meier estimate (O'Quigley & Xu, 2001).
The weighted version of the new measure R_{sch}^2 is then written as follows:

$$\begin{aligned} R_{sch}^2(\beta) &= \frac{\sum_{i=1}^{n} \delta_i W(t_i) \{\beta' r_{sch_i}(0)\}^2 - \sum_{i=1}^{n} \delta_i W(t_i) \{\beta' r_{sch_i}(\beta)\}^2}{\sum_{i=1}^{n} \delta_i W(t_i) \{\beta' r_{sch_i}(0)\}^2} \\ &= 1 - \frac{\sum_{i=1}^{n} \delta_i W(t_i) \{\beta' r_{sch_i}(\beta)\}^2}{\sum_{i=1}^{n} \delta_i W(t_i) \{\beta' r_{sch_i}(0)\}^2} \end{aligned}$$

The weights $W(t_i)$ are the height of the step of the marginal Kaplan-Meier survival curve at time t_i; $r_{sch_i}(\beta)$ is the Schoenfeld residual for the covariate model and $r_{sch_i}(0)$ the residual for the null model.
$R_{sch}^2(\beta)$ is a consistent estimate for $\Omega^2(\beta)$, which is a measure of explained variation for $X|t$ (O'Quigley & Xu, 2001). The expected minimum is 0 and the maximum is 1 while $R_{sch}^2(\beta)$ is increasing with β. Another advantage of this formulation is that asymptotically, decomposition into residual sums of squares is possible, as for linear models:

$$SST \overset{asympt.}{=} SSR + SSM$$

with

$$\begin{aligned} SST &= \sum_{i=1}^{n} \delta_i W(t_i) \left\{\hat{\beta}' r_{sch_i}(0)\right\}^2 \\ SSR &= \sum_{i=1}^{n} \delta_i W(t_i) \left\{\hat{\beta}' r_{sch_i}(\hat{\beta})\right\}^2 \\ SSM &= \sum_{i=1}^{n} \delta_i W(t_i) \left\{\hat{\beta}' E_{\hat{\beta}}(X_{t_i}) - \hat{\beta}' E_0(X|t_i)\right\}^2 \end{aligned}$$

Apart from these properties, this formulation allows for an easy extension to nested, stratified and time-varying models. An overview of possible extensions and proofs are given in Xu (1996), Xu & O'Quigley (2001) and Xu & Adak (2002).

For comparison reasons, a new variant based on Schoenfeld residuals is introduced which is generally constructed like the measures based on cumulative hazards while the weighting is kept. The mean absolute difference between the residuals of the two models is therefore calculated, weighted and divided by the mean of weighted absolute residuals of the null model. The new measure is written:

$$R^2_{sch.k}(\beta) = \frac{\frac{1}{n}\sum_{i=1}^{n} \delta_i W(t_i) \left|\beta' r_{sch_i}(0) - \beta' r_{sch_i}(\beta)\right|}{\frac{1}{n}\sum_{i=1}^{n} \delta_i W(t_i) \left|\beta' r_{sch_i}(0)\right|}$$
$$= \frac{\sum_{i=1}^{n} \delta_i W(t_i) \left|\beta' r_{sch_i}(0) - \beta' r_{sch_i}(\beta)\right|}{\sum_{i=1}^{n} \delta_i W(t_i) \left|\beta' r_{sch_i}(0)\right|}$$

Being aware that some desirable properties of $R^2_{sch}(\beta)$ are lost by this new formulation, this measure is mainly introduced to draw a comparison with $K_{m.norm}$ and $K_{d.norm}$.

4 Simulation studies

4.1 Data simulation

In simulation studies, the presented measures were analysed with respect to their dependence on coefficients β, distribution of the covariate X and censoring percentage in the data. In addition, relations between the outcomes of the different measures were investigated.
Computation of the measures was based on data sets consisting of 1000 observations with exponentially distributed failure time t_f and expectation $E(t_f|X,\beta) = 1/\exp(\beta' X)$. Uninformative censoring was added, comparable to clinical studies where patients enter continuously over time and the study is stopped after a predefined maximum observation time or after a certain number of events is reached. Therefore, a uniformly distributed censoring time was created for each subject, $t_c \sim U(0,\tau)$. The observation time is taken as the minimum of t_f and t_c, the status indicator is set to 1 for $t_f \leq t_c$ and zero otherwise. The upper limit τ was varied to obtain censoring percentages of 0% and approximately 10%, 25%, 50% and 80%. Hence, the influence of increasing censoring can be investigated.
Distributions of X were initially chosen as binary with $p = 0.5$, uniform $X \sim U(0,\sqrt{3})$ and normal $X \sim N(0, 0.25)$ as these three distributions have the same variance although they differ in expectation. Additionally, more standard distributions were chosen, $X \sim N(0,1)$ and $X \sim U(0,1)$. The former of which results in higher variance of X with values centred at 0 while the latter will have lower variance than the binary covariate. The influence of the distribution of X can therefore be observed.
Each value of X was then associated with different coefficients chosen from the set $\beta \in \{0.5, 1, 2, 3\}$ to yield failure times.
For each of the five covariate distributions, 200 data sets of 1000 observations along with one failure time and four different censoring times for each value of β were generated. Hence, each measure was calculated 200 times for each setting defined by covariate distribution, value of β and censoring percentage.

4.2 Criteria investigation

All criteria were initially calculated for the data sets with binary X. The obtained criteria are plotted against the according coefficients β in figure 1. Lines are drawn between the means at each value of β with different styles for different censoring percentages in the data. Correlations between β and the values of all the criteria are high. The maximum correlation is 0.995 and is reached by R^2_{sch} for data without censoring whereas the minimum is 0.910 and is obtained for V_W in the data with 80% censoring.
As can be seen in figure 1, all criteria grow with increasing values of β, which

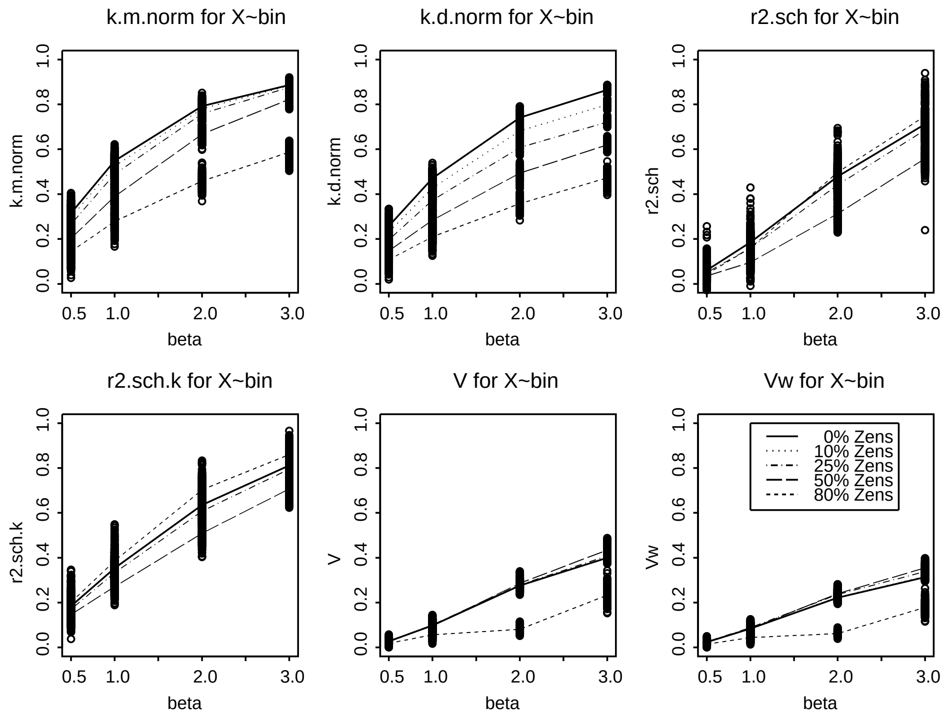

Figure 1. Criteria over different values of coefficient β and varying censoring percentages for a binary covariate X. The points display the true values; lines are drawn between the criteria's means for each coefficient. The lines represent the mean results for different censoring percentages in the data as indicated in the legend.

is desirable as covariates with high coefficients are more important by means of prediction. The strength of the increase is, to some extend for all criteria, dependent on the censoring percentage in the data. The measures based on the variance of the survival curves are less affected by censoring providing this is not extreme. For high censoring (80%) the factor's contribution to the model can only be detected for high coefficients, here $\beta = 3$ (figure 1). The measures based on martingale and deviance residuals are more obviously affected by censoring than the others. The strong monotonic decrease is due to less extreme values of the cumulative hazard in presence of censoring. Hence, residuals are less extreme especially for the covariate model, and the difference between the two models decreases. The Schoenfeld residual measures yield the lowest values for 50% of censoring and, in contrast to the other criteria, highest values for the data with 80% censoring. But they are generally less affected by censoring than the martingale and deviance residual measures. However, the values of these criteria have higher variance.

When comparing the absolute values of all criteria, wide differences occur (table 1). For data without censoring and coefficient $\beta = 3$, the martingale residual measure yields an average of 0.886 whereas the mean of V_W is 0.314.

Table 1. Mean values of the criteria for 200 simulated data sets per coefficient β without censoring and binary X.

Measure	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 3$
K.m.norm	0.315	0.548	0.792	0.886
K.d.norm	0.257	0.471	0.741	0.864
R2.sch	0.060	0.185	0.479	0.712
R2.sch.k	0.184	0.354	0.635	0.811
V	0.027	0.098	0.275	0.400
Vw	0.024	0.084	0.222	0.314

The Schoenfeld residual measures yield values that range between martingale and survival measures. Therefore the criteria are grouped and a general ranking can be established which in most cases holds for all tested covariate distributions:

$$K_{m.norm}, K_{d.norm} \geq R^2_{sch}, R^2 sch.k \geq V, V_W$$

In addition, the shape of the trend of the criteria over β can be judged. For the measures based on martingale or deviance residuals, the trend over β is more logarithmic whereas the other measures show more linear dependencies. This is due to the expected maximum value of 1. It will be seen in other simulation data that as soon as the measures approach 1, the curve flattens for all the criteria.
On the other hand, the criterion R^2_{sch} in some cases is slightly negative for $\beta = 0.5$. In these cases the measure presumes that the effect of the covariate is negligibly small i.e. the null model is better than the covariate model. In fact, $\beta = 0.5$ is a small coefficient for a binary covariate. The associated relative risk would be $rr = \exp(\beta) = 1.65$. And most of the other criteria also yield some values near zero.

The same analysis was then carried out on the continuous covariate distributions. First choice was X~$U(0, \sqrt{3})$ as this distribution has the same variance as that of a binary variable with $p = 0.5$ and differs only slightly in its expectation. The results are displayed in figure 2. As can be seen, the criteria behave much like those calculated for the data sets with a binary covariate. Only for the measures V and V_W the increase with β in the high censoring data, 80%, is more linear than for binary data. In addition, there are more extreme values within R^2_{sch} and $R^2_{sch.k}$. Standard deviations of these criteria increase for high censoring percentages more than for the others.

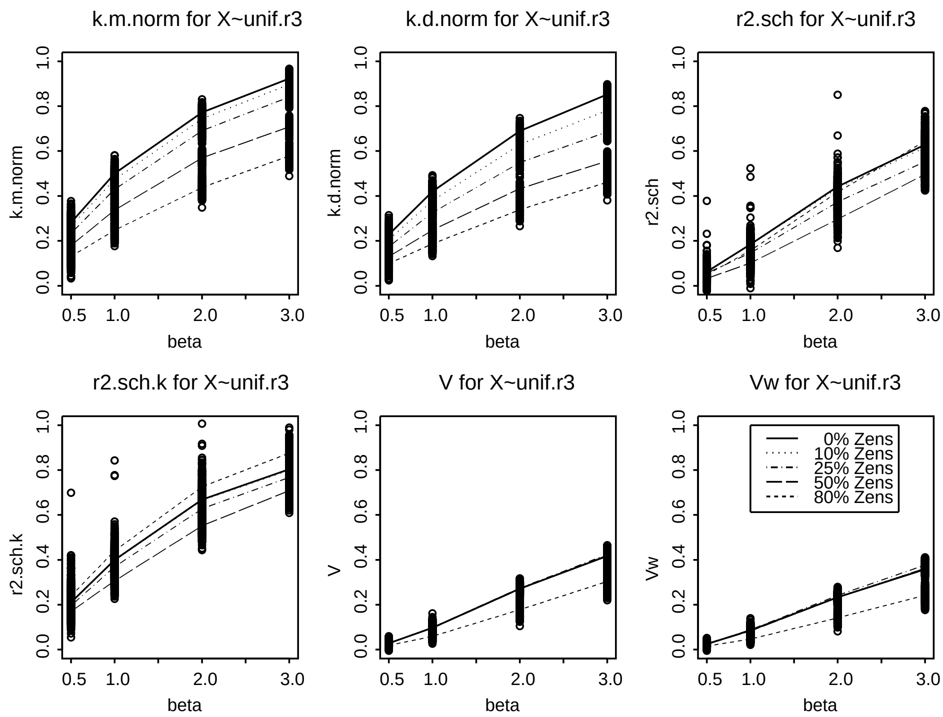

Figure 2. Criteria over different values of coefficient β and varying censoring percentages for a uniform covariate distribution $X\sim(0, \sqrt{3})$. The points display the true values; lines are drawn between the criteria's means for each coefficient. The lines represent the mean results for different censoring percentages in the data as indicated in the legend.

The next data analysed was that with normally distributed covariates, $X\sim N(0, 0.25)$. The values of X are centred at zero but have the same variance as the other distributions that have already been discussed. The centring of X at zero leads to lower values in all the criteria as can be seen in figure 3. The measures V and V_W hardly yield more than 0.2 and several values of R^2_{sch} are again slightly negative. X in this setting is therefore not a strong factor.

The data with $X\sim U(0, 1)$ was then analysed with respect to the model fit criteria. The expectation of X is the same as for binary X although the variance is lower. The results are very similar to those for the data with $X\sim N(0, 0.25)$. Hence, no separate plot is displayed. Again, values of V and V_W are very low and rarely exceed 0.2. The standard deviations of the Schoenfeld residual measures increase with the censoring percentage in the data.

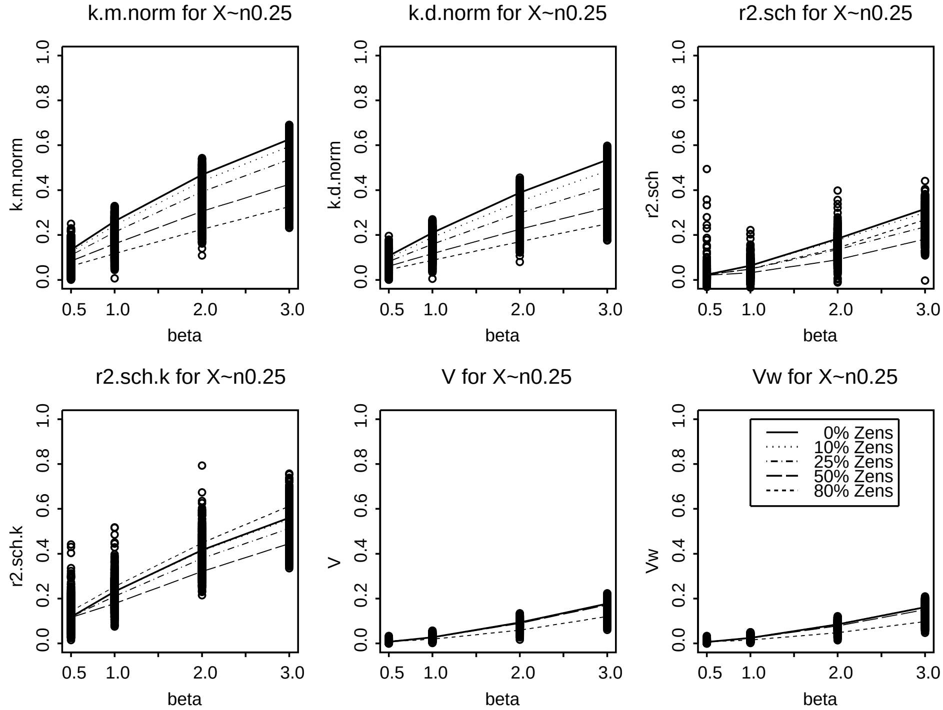

Figure 3. Criteria over different values of coefficient β and varying censoring percentages for a uniform covariate distribution X~$N(0, 0.25)$. The points display the true values; lines are drawn between the criteria's means for each coefficient. The lines represent the mean results for different censoring percentages in the data as indicated in the legend.

The last analysed data sets are those with a standard normally distributed covariate. The covariate values are centred at zero but have variance four times as high as the binary covariate. The high variance allows for high values of X. In combination with high coefficients β this leads to high values in criteria judging the goodness of fit as can be seen in figure 4.

Here, all curves have more or less logarithmic shapes. The values of all the criteria are much higher than in the cases discussed before. Especially $K_{m.norm}$ and $K_{d.norm}$ exceed the expected upper limit of 1 for high coefficients. As mentioned before, cumulative hazards increase for strong factors and lead to extreme residuals such that the absolute difference between the residuals in the covariate model and those from the null model is higher than the absolute residuals calculated for the null model. Therefore an extreme improvement is obtained. The same observation is made for a few values of $R^2_{sch.k}$. The question arises whether this setting is a realistic situation. If so, the fact that the limit of 1 is exceeded requires correction of these criteria as the property of interpretation as a percentage of explained variation is lost otherwise. On the other hand, the values of V and V_W are very low for

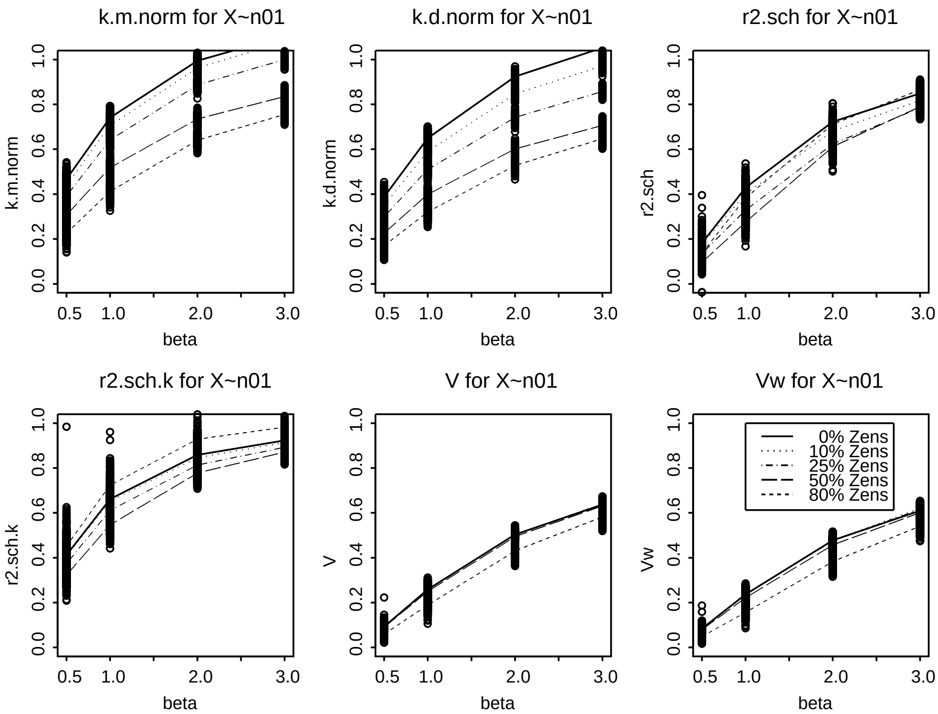

Figure 4. Criteria over different values of coefficient β and varying censoring percentages for a uniform covariate distribution X-$N(0,1)$. The points display the true values; lines are drawn between the criteria's means for each coefficient. The lines represent the mean results for different censoring percentages in the data as indicated in the legend.

all other distributions of X and need further examination in case the other settings are more realistic.

Summary

In summary, it is evident that all the criteria strongly depend on the coefficient associated with the covariate. The measures $K_{m.norm}$ and $K_{d.norm}$ decrease constantly with increasing censoring percentages and meanwhile it has been established in this piece of work that they may exceed the expected maximum of 1 for extreme settings.
The Schoenfeld residual based measures return the lowest values for 50% censoring in the data and are often highest for 80%. They have generally higher variance than the other measures. For low coefficients, R^2_{sch} may be slightly negative, which is an indication for a very weak covariate. In these cases, the null model is preferred over the covariate model. $R^2_{sch.k}$, as well as $K_{m.norm}$ and $K_{d.norm}$, exceeds the desired maximum value of 1 in extreme settings. Therefore, these criteria cannot be interpreted as a percentage of explained variation.
V and Vw only decrease for high censoring (80%). They are otherwise unaffected by censoring but the values are generally low.
All measures highly depend on the true underlying covariate distribution on the basis of which the survival times are created. None of the criteria are affected by rescaling of the covariate in the criteria calculating process, as they all involve the covariate only in combination with the associated coefficient β. Rescaling of the covariate only results in a different value of β during estimation.

4.3 Relations between the different criteria

As all of the presented criteria are supposed to measure the goodness-of-fit of a model, the next idea was to examine the relations between them. Analytical descriptions of these are difficult, as the criteria building processes differ in number of the terms in the sums and weighting functions are different.
The criteria resulting from the 800 simulation data sets per covariate distribution without censoring (200 per value of β) were therefore plotted against each other. Strong relationships occurred. When comparing the plots for all covariate distributions, the relationships seemed very similar for all continuous covariate distributions. When analysing all criteria for continuous X together (i.e. 3200 values per criterion), correlations ranged between 0.9318 and 0.9980 whereas for binary X, correlations ranged between 0.9630 and 0.9997. Summing all results of the five distributions together changed the range of correlations to (0.9313, 0.9970). The plot, however, showed a slightly different trend for the criteria resulting from simulation data with binary X for several combinations. Hence, it was decided to keep the distinction between continuous and binary covariate distributions.
In figure 5, the criteria resulting from simulation data without censoring and continuous distributions of X are plotted against each other. The strong relationship between the criteria is obvious. The gaps that occur in the plot can be explained by the discrete values of β. Although the trend is obvious for all relationships, those which involve $R^2_{sch.k}$ show higher variance.
The plot for the data with binary X shows similar trends and comparably strong relationships between the criteria. Therefore, it is not displayed separately. The regions for different values of β, however, are more distinguishable except for the relation between $K_{m.norm}$ and $K_{d.norm}$, which is shaped like the relation in figure 5. Again, the relations to $R^2_{sch.k}$ have higher variance.

As no analytical answer to the question of the true form of the relationship is available yet, *fractional polynomials* were applied to each pair of criteria. In this way, a first impression of the relationships and the influence of censoring percentage, covariate distribution and coefficient on these can be obtained. Fractional polynomials are defined as follows (Royston & Altman, 1994):

$$FP(x,p) = b_0 + \sum_{j=1}^{m} b_j x^{(p_j)}$$

Hence, a polynomial of degree m with exponents p is fit to describe the form of a trend. In practice, a maximum degree of $m = 2$ and exponents chosen from the set $p \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, with $x^{(0)} = \ln(x)$, is sufficient for describing most trends. Each exponent can be chosen more than once. In this case, the first term will be x^p and the second is defined

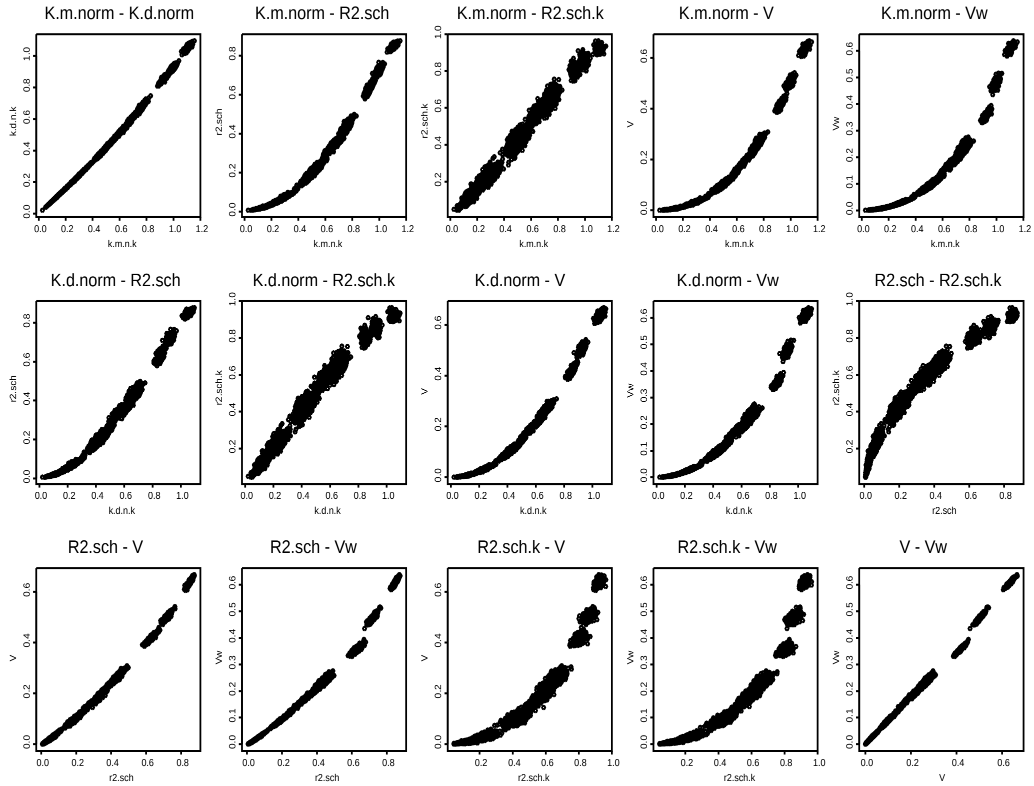

Figure 5. Criteria for continuous distributions of X for data without censoring plotted against each other show strong relationships.

as $\ln(x)x^{(p)}$. Consequently, a fractional polynomial of degree $m = 2$ with exponents $p = (0, 3)$ for example is written:

$$FP(x, p = (0, 3)) = b_0 + b_1 \ln(x) + x^3$$

The optimal fractional polynomial is found stepwise starting from the most complex model with $m = m_{max}$. This is compared to the best model of degree $m = m_{max} - 1$. The procedure stops as soon as the deletion of a term results in a significant change in deviance.

The goodness-of-fit of a fractional polynomial model is measured by residual deviance, which is also the measure of choice for judging the goodness-of-fit of generalised linear models. A good fit is achieved if the residual deviance is small. The residual deviances, along with the corresponding number of residual degrees of freedom, are displayed in table 2 for the two data sets. Compared to the number of observations and remaining residual degrees of freedom, the residual deviance is generally small. This can be seen in table 2 and therefore gives an indication for a good fit. As already pointed out in the discussion of the plots, the highest variance occurs for relations to

Table 2. Residual deviances resulting from fractional polynomial fits between the criteria calculated for continuous distributions of X and binary X respectively.

Model	Residual deviance X~ cont. Residual $df = 3197$	Residual deviance X~ bin. Residual $df = 797$
$K_{d.norm} \sim K_{m.norm}$	0.1531	0.0857
$R^2_{sch} \sim K_{m.norm}$	0.5202	0.3154
$R^2_{sch.k} \sim K_{m.norm}$	2.4110	0.4277
$V \sim K_{m.norm}$	0.2219	0.0916
$V_W \sim K_{m.norm}$	0.4758	0.0436
$R^2_{sch} \sim K_{d.norm}$	0.7038	0.0359
$R^2_{sch.k} \sim K_{d.norm}$	2.5195	0.1514
$V \sim K_{d.norm}$	0.2479	0.0055
$V_W \sim K_{d.norm}$	0.5925	0.0017
$R^2_{sch} \sim R^2_{sch.k}$	1.6436	0.2189
$R^2_{sch} \sim V$	0.1793	0.0259
$R^2_{sch} \sim V_W$	0.2915	0.0165
$V \sim R^2_{sch.k}$	1.0901	0.0690
$V_W \sim R^2_{sch.k}$	1.3063	0.0374
$V \sim V_W$	0.1859	0.0011

$R^2_{sch.k}$, which is proved by higher residual deviances. The relation, however, is strong.

Variance in the plot increased with censoring in the data, especially for relations to the Schoenfeld residual measures. The trend between the two criteria based on cumulative hazards and that between the measures V and V_W remained practically unchanged, which indicates similar dependencies on censoring for these criteria. The reason for the high variance is obviously the higher variance that has been observed for the criteria R^2_{sch} and $R^2_{sch.k}$. $K_{m.norm}$ and $K_{d.norm}$, however, are strongly dependent on censoring. Therefore, the relation to the other criteria is increasingly compressed, although the general shape of the trend is kept.
As for the data with binary X, the detection of a trend for relations involving V or V_W is difficult for 80% censored observations per data set. As already mentioned in the discussion of figure 1, these criteria only detect the covariate when it is combined with a high coefficient. The trend over β is not smooth. Consequently, the relation to the other criteria is not smooth either.

4.4 Discussion of simulation results

The presented criteria judge the goodness-of-fit of a survival model with respect to different outcomes, as there are cumulative hazard; prediction of covariates, and variance of survival curves. All depend strongly on the associated coefficient and the distribution of the true underlying covariate. However, they differ strongly in value. While the measures V and V_W are generally very small, the criteria based on cumulative hazard tend to exceed the expected maximum value of 1 for strong factors in data sets with very small censoring percentages. The latter measures are the only ones that obviously decrease monotonically with increasing censoring in the data. $R^2_{sch.k}$ also exceeds the desired maximum value of 1, and it therefore does not allow for interpretation as a percentage. R^2_{sch} for weak factors occasionally yields slightly negative values, although its minimum in expectation is 0. In addition, the measures based on Schoenfeld residuals have higher variance than all the other criteria presented here. Consequently, all of the measures suffer drawbacks.
However, strong relations between all of the criteria could be detected. When describing these, data with binary X had to be distinguished from data with continuous X. For further specification of the trends the censoring percentage has to be taken into account. This is especially the case for relations including $K_{m.norm}$ and $K_{d.norm}$. Once this is realised, the successful calculation of at least one of the measures should allow for the subsequent derivation of all remaining measures. The precision of covariate prediction is therefore directly related to the gain of explained variation in the survival curves and the precision of cumulative hazards.

5 Application to stomach cancer data

The data analysed originate from a clinical study from *Chirurgische Klinik der TU München* during the years 1987 and 1996. 295 patients with stomach cancer were analysed with respect to their survival. The maximal individual follow-up time is 11 years. The censoring percentage in the data is 63%. In earlier analyses with less data (Stark, 1997), a dichotomised version of the percentage of seized lymph nodes (NODOS.PR) was identified as strongest prognostic factor. The optimal version of NODOS.PR is now recalculated for the new data and analysed with respect to its contribution to the goodness-of-fit of the model by means of the presented criteria.

For a complete analysis, assumptions for the application of a Cox model must be checked first. These include linearity (if necessary, combined with the finding of the optimal functional form) of the covariate and proportional hazards. Initially, a varying coefficient model (Hastie & Tibshirani, 1993) is fit using a spline of NODOS.PR as covariate to check for linearity. As can

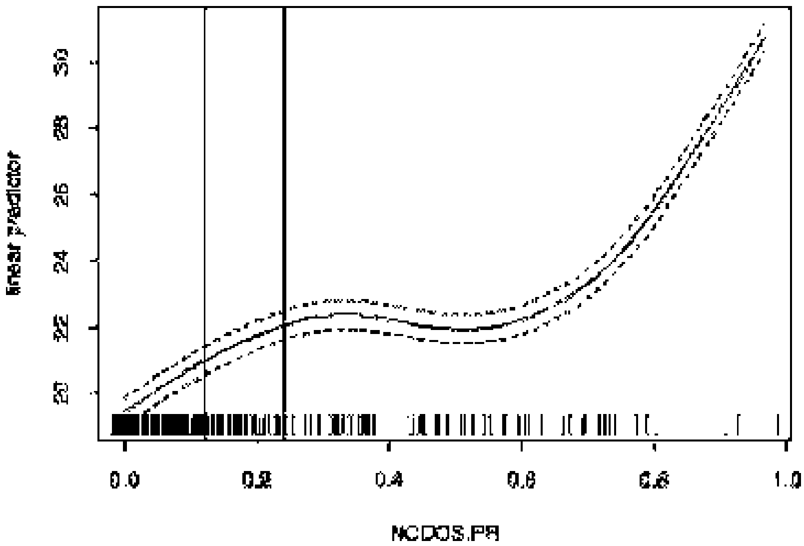

Figure 6. The varying coefficients spline plot for the continuous covariate NODOS.PR shows that the linearity assumption is violated. No linear trend is possible within the 95% confidence region (dashed lines). Vertical lines indicate possible positions for dichotomisation.

be seen in figure 6, no linear effect is possible within the 95% confidence region (dashed lines) and a dichotomised version of NODOS.PR is preferable which distinguishes between a low risk and a high risk group.
Another indication for dichotomisation was found in the check of the functional form of NODOS.PR. This graphical check is established by plotting martingale residuals of the null model against the values of NODOS.PR (Therneau & Grambsch, 2000). The smoothed fit is monotonic and logistically shaped (figure 7), which is an indication for a threshold, and a dichotomisation is proposed.
The optimal cut point was found by simultaneously testing all data points in NODOS.PR, that guarantee at least ten individuals per group, as candidate split points. Test statistic is the logrank statistic, which is adjusted for multiple testing according to Lausen & Schumacher (1992). The optimal cut point is found as the one with minimum p-value. Here, the two minimal p-values have been picked. The according cut points lead to two different binary covariates that were tested with respect to their contribution to the model fit. The optimal cut point was found at 0.1212, the next best point for dichotomisation is at 0.2326. Both log rank statistics are highly significant even after adjusting the p-values for multiple testing. The new covariates are named *nod.122* and *nod.24*.
Next, the assumption of proportional hazards was checked using the method proposed by Grambsch & Therneau (1994), which tests for correlation between time, or transforms of time, and scaled Schoenfeld residuals. Here,

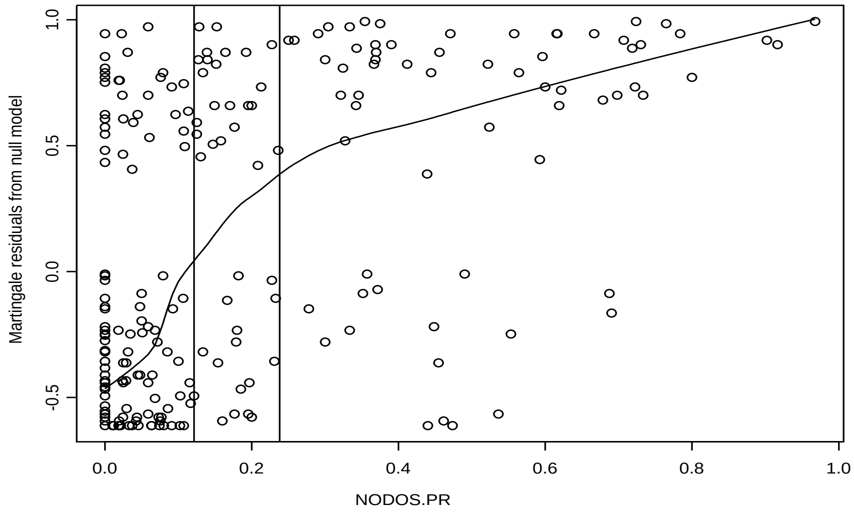

Figure 7. Martingale residuals obtained from the null model plotted against the covariate values indicate the functional form of the covariate. Here, a dichotomisation is chosen. Proposed cut points are indicated as vertical lines.

correlations with time, ranks of time and logarithmic time were tried. No violation of the model assumption could be found.

The two univariate models were therefore analysed with respect to the presented goodness-of-fit criteria. The coefficients in the two models were estimated as $\beta_{nod.122} = 2.04$ and $\beta_{nod.24} = 1.95$. The results for all goodness-of-fit criteria are displayed in table 3.

Table 3. Goodness-of-fit criteria for two univariate models including a binary variable for NODOS.PR and means of criteria obtained from simulation studies with β=2 combined with 50% and 80% censoring in the data.

Criteria	**Model with** *nod.122*	**Model with** *nod.24*	**Simulation with** β=2, Zens=50%	**Simulation with** β=2, Zens=50%
$K_{m.norm}$	0.5897	0.4531	0.6663	0.4571
$K_{d.norm}$	0.4482	0.3541	0.4929	0.3569
R^2_{sch}	0.3242	0.2733	0.3127	0.4974
$R^2_{sch.k}$	0.5529	0.46687	0.5085	0.7019
V	0.2595	0.2296	0.2850	0.0806
V_W	0.2140	0.1982	0.2389	0.0617

The factor *nod.122* provides higher values in all the criteria. Therefore, it should be preferred over *nod.24*. As in the simulation study, the values of the different criteria show high discrepancies in both models.
The results for the model built with the factor *nod.122* are interpreted as follows: approximately 59% of absolute martingale residual deviation can be explained by introducing *nod.122* into the model, while 32% of the variation of Schoenfeld residuals can be explained and the variation in the survival curves is reduced by 26%. These values are generally high and *nod.122* is considered as an important factor for prediction.
Additionally, all criteria for the model with *nod.122* were compared to the means calculated in our simulation study in the previous section for 50% and 80% censoring in data with a binary covariate associated with $\beta = 2$, as these settings are closest to the real data. The simulation study showed a constant decrease with increasing censoring percentage only for $K_{m.norm}$ and $K_{d.norm}$ (figure 1). Simulation data with 63% censoring would therefore be expected to yield criteria values between those calculated for 50% and 80% censoring. Although the simulation data are optimally created, the values of these criteria obtained for the real data are close to the point where simulation data would be expected to yield values. The other criteria's values are still between the simulation means but closer to those obtained for 50% censoring. For these criteria, the exact analytical dependence on censoring percentage is not available, yet. Therefore, it is difficult to establish, where the expected values should be in an optimal setting. But the results from the real data seem not to clearly contradict the simulation results. Taking into account that the simulation is optimally created for a univariate model and the resting variance in the data is random and therefore cannot be explained, the model for the real data (based on *nod.122*) seems to be very good.

6 Conclusion

Different criteria have been presented to measure the goodness-of-fit in survival analysis. They are all obtained through different procedures and show high discrepancies in value when calculated simultaneously for the same data. All of them have drawbacks. The measures $K_{m.norm}$, $K_{d.norm}$ and $R^2_{sch.k}$ tend to exceed 1 in extreme settings and can therefore not generally be interpreted as a percentage of explained variation. Correction would be needed. It is difficult, however, as the problem is in the structure of martingale residuals and the use of absolute distances. The measures V and V_W are generally very low, whereas R^2_{sch} has high variance and in some cases of low associated coefficients, yields slightly negative values, which indicates a weak factor. The latter measure, however, has been extensively studied and allows for interpretation as a residual sum of squares in the classical sense; an approximated decomposition into sums of squares and easy extensions to different other settings. It is therefore currently recommended for use. However, the values of the different criteria indicate that they are strongly connected to one another. All the criteria are of interest, and the analytical derivation of the relations between them is the aim of further research, as well as the formulation of a model selection procedure based on an appropriate measure of explained variation for survival data.

References

[1] Altman, D.G., De Stavola, B.L., Love, S.B., Stepniewska, K.A. (1995). Review of survival analyses published in cancer journals. *Br J Cancer*, **72**, 511-518

[2] Berger, U., Schaefer, J., Ulm, K. (2003). Dynamic Cox modelling based on fractionall polynomials: time-variations in gastric cancer prognosis. *Statistics in Medicine*, **22**, 1163-1180

[3] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*,**30**, 89-100

[4] Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Royal Stat. Soc. B*, **34**, 187-220

[5] Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276

[6] Grambsch, P.M., Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*,**81**, 515-526

[7] Hastie, T., Tibshirani, R. (1993). Varying-coefficient models. *J. Royal Stat. Soc. B*, **55**, 757-796

[8] Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481

[9] Kendall, M. (1975). Rank correlation methods. 4th Editio, London and High Wycombe, Griffin.

[10] Kvalseth, T.O. (1985). Cautionary note about R^2. *The American Statistician*, **39**, 279-285

[11] Lausen, B., Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics*, **48**, 73-85

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

[13] O'Quigley, J., Xu, R. (2001). Explained variation in proportional hazards regression. *Handbook of Statistics in Clinical Oncology.* Ed.: Crowley. Marcel Dekker, Inc.2001, 397-410

[14] Royston, P., Altman, D. (1994). Regression using fractional polynomials of continuous covariate: parsimonious parametric modeling. *Applied Statistics*, **43**, 429-467

[15] Schemper, M. (1990). The explained variation in proportional hazards regression. *Biometrika*, **77**, 216-218

[16] Schemper, M., Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics*, **56**, 249-255

[17] Schemper, M., Smith, M.S. (1996). A note on quantifying follow-up in studies of failure time. *Control Clin Trials*, **17**, 343-346

[18] Schoenfeld, D.A. (1982) Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241

[19] Stark, M. (1997). Beurteilungskriterien fuer die Guete von Modellen zur Analyse von Ueberlebenszeiten. *Logos Verlag*, Berlin

[20] Therneau, T., Grambsch, P. (2000). Statistics for Biology and Health. Modeling Survival Data - Extending the Cox Model. Springer-Verlag New York Berlin Heidelberg

[21] Therneau, T., Grambsch, P. ,Fleming, T. (1990). Martingale based residuals for survival models. *Biometrika*, 77, 147-160

[22] Xu, R. (1996) Inference for the proportional hazards model. *PhD thesis of University of California, San Diego*

[23] Xu, R., Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, 58, 305-315