Predictive Power of the Cox Model: Measuring the Regression Effect as a Function of Explained and Unexplained Variation

Ronghui Xu[1] and John O'Quigley[2]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

[2]Department of Mathematics, University of California at San Diego, La Jolla, CA 92093

[1]rxu@jimmy.harvard.edu [2]oquigley@math.ucsd.edu

SUMMARY: An R^2 type index of the predictive power of the Cox model has recently been described (O'Quigley and Xu, 2001). Under broad conditions the index can also be viewed as a measure of explained variation and builds on an earlier proposal (O'Quigley and Flandre, 1994). One interesting and practically useful property is the ability to accommodate time dependent covariates. In this paper we study more deeply the motivation and interpretation of the index as well as showing, via practical examples, how the R^2 index can throw light on the complex interrelations between the risk factors and prognosis in cancer studies. The index can be shown to: (1) take a value of zero in the absence of regression effect, (2) increase to the value 1 when regression effect tends to infinity and prediction of the ranks becomes deterministic, (3) has the property that increasing values of its population counterpart Ω^2 correspond to increasing predictability of the ranks of the survival times, (4) remains invariant to linear transformation of the covariates and to increasing monotonic transformations of time, (5) enjoys a concrete interpretation in terms of sums of squares decompositions and the ability to express Ω^2 as a proportion of explained variation. Furthermore Ω^2 can be shown not to be affected by an independent censorship. Through simulation studies we compare the index with other proposals in the literature.

Key words: Explained variation; Kaplan-Meier estimator; Predictability; Prognostic index; Proportional hazards regression; Schoenfeld residuals.

1 Introduction

1.1 Motivating clinical problem

In a study of 2174 breast cancer patients, followed over a period of 15 years at the Institut Curie in Paris, France, a large number of potential and known prognostic factors were recorded. Detailed analyses of these data have been the subject of a number of communications and we focus here on a limited analysis on a subset of prognostic factors, identified as having some prognostic importance. These factors were: 1) age at diagnosis, 2) histology grade, 3) stage, 4) progesterone receptor status, and 5) tumor size. In addition to the usual model fitting and diagnostic tools, it seems desirable to be able to present summary measures estimating the percentage of explained variation. Making precise the notion of explained variation in the context of proportional hazards regression, in which we allow inference to be invariant to unspecified monotonically increasing transformations on time, is not immediate but a suitable measure would reflect the relative importance of the covariates.

We would like to be able to say, for example, that stage explains some 20% of survival but that, once we have taken account of progesterone status, age, and grade, then this figures drops to 5%. Or that adding tumor size to a model in which the main prognostic factors are already included then the explained variation increases, say, a negligible amount, specifically from 32% to 33%. Or, given that a suitable variable indicates predictability, then to what extent do we lose (or gain), in terms of these percentages, by recoding the continuous prognostic variable, age at diagnosis, into discrete classes on the basis of cutpoints.

For a measure to be able to deal with the above requirements it would be required to have certain properties: the above percentages should be meaningful and directly related to predictability of the failure ranks, absence of effect should translate as 0%, perfect prediction of the survival ranks should translate as 100% and intermediate values should be interpretable.

1.2 A brief overview

For the proportional hazards model correlation measures were first suggested by Maddala (1983) although the measure depends heavily on censoring. Kent and O'Quigley (1988) developed a measure based on the Kullback-Leibler information gain and this could be interpreted as the proportion of randomness explained in the observed survival times by the covariates. The principle difficulty in Kent and O'Quigley's measure was its complexity of calculation although a very simple approximation was suggested and appeared to work well. The Kent and O'Quigley measure was not able to accomodate time-dependent covariates. Xu and O'Quigley (1999) developed a similar measure based on information gain, using the conditional distribution of the covariates given the failure times. The measure accomodates time-dependent covariates, and is computable using standard softwares for fitting the Cox model. Some current work is being carried out on the measures of explained randomness, and will be available in a separate manuscript. Schemper (1990, 1994) introduced the concept of individual survival curves for each subject, with the model and without the model. Interpretation is difficult. As with the Maddala measure the Schemper measures depend on censoring, even when the censoring mechanism is completely independent of the failure mechanism (O'Quigley et al., 1999). Korn and Simon (1990) suggested a class of potential functionals of interest, such as the conditional median, and evaluated the explained variation via an appropriate distance measuring the ratio of average dispersions with the model to those without a model. Their measures are not invariant to time transformation nor could they accommodate time-dependent covariates. Korn and Simon (1990) also suggested an approach to calculate Somer's D; this in general requires numerical integration and depends on a common censoring time for all survival times. Schemper and Kaider (1997) proposed to estimate the correlation coefficient between failure rankings and the covariates via multiply imputing the censored failure times. Other measures have also been proposed in the literature, but it is not our intention here to give a complete review of them.

O'Quigley and Xu (2001) described a measure of explained variation for proportional hazards regression, which is a further development of O'Quigley and Flandre (1994). The measure is recalled in the next section. In section 3 we study the statistical properties of the measure. These properties are particularly attractive from the viewpoint of interpretation, in particular the ability to obtain a sum of squares decomposition analogous to that for the linear case, an expression in terms of explained variation and the observation that increasing values of the measure directly translate increasing predictability of the survival ranks. In section 4 via simulations we compare this measure with some of the measures mentioned above. Following that we discuss the partial coefficients and the extensions of the measure to other relative risk models.

Finally, an illustration of the practical usefulness of the measure in analyzing survival data under the proportional hazards model is given in section 6. This is done via three illustrative examples: breast cancer, gastric cancer and multiple myeloma. In these studies we are interested in questions such as how much additional prognostic value is contained in some explanatory variable once others have been included in the model, to what extent the observed failure rankings can be explained by the totality of the explanatory variables, to what extent modelling time-dependent effects improves prediction, and to what extent predictive power is diminished, if at all, by recoding a continuous covariate as a discrete dichotomy or trichotomy. The example of multiple myeloma is interesting in that, even using all available molecular and genetic markers, we fail to account for over 80% of the variability in the survival rankings. For this example it is also interesting that biomarkers alone would appear to capture all the predictive capacity of the traditional Durie-Salmon staging system.

2 Model and R^2 measure

2.1 Model and notation

In a survival study denote T the potential failure time, and C the potential censoring time. Let $X = \min(T, C)$, $\delta = I(T \leq C)$ where $I(\cdot)$ is the indicator function, and $Y(t) = I(X \geq t)$. Associated with T is the vector of possibly time-dependent covariates $Z(t)$. For our mathematical development, assume $(T_i, C_i, Z_i(\cdot))$, $i = 1, 2, ..., n$, to be a random sample from the distribution of $(T, C, Z(\cdot))$. We will also use the counting process notation: let $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $\bar{N}(t) = \sum_1^n N_i(t)$. For most of this work we assume a conditional independent censorship model. However, under the stronger assumption of independent censorship, where C is assumed to be independent of T and Z, we obtain further properties and interpretation as described in Section 3.

The Cox (1972) proportional hazards model assumes that the conditional hazard function

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta' Z(t)\}, \tag{2.1}$$

where $\lambda_0(t)$ is an unknown "baseline" hazard, and β is the relative risk parameter. Denote

$$\pi_i(\beta, t) = \frac{Y_i(t)\exp\{\beta' Z_i(t)\}}{\sum_{j=1}^n Y_j(t)\exp\{\beta' Z_j(t)\}}; \tag{2.2}$$

it is the conditional probability of subject i being chosen to fail, given all the individuals at risk at time t and that one failure occurs. The product of the π's over the observed failure times gives Cox's (1972, 1975) partial likelihood. When $\beta = 0$, $\{\pi_i(0,t)\}_{i=1}^n$ is simply the empirical distribution, assigning equal weight to each sample subject in the risk set. Denote the expectation of a variable with respect to $\{\pi_i(\beta,t)\}_{i=1}^n$ by $\mathcal{E}_\beta(\cdot|t)$. In particular

$$\mathcal{E}_\beta(Z|t) = \sum_{j=1}^n Z_j(t)\pi_j(\beta,t) = \frac{\sum_{j=1}^n Y_j(t) Z_j(t)\exp\{\beta' Z_j(t)\}}{\sum_{j=1}^n Y_j(t)\exp\{\beta' Z_j(t)\}} \tag{2.3}$$

is the expectation of $Z(t)$ with respect to $\{\pi_i(\beta,t)\}_i$, and

$$r_i(\hat\beta) = Z_i(X_i) - \mathcal{E}_{\hat\beta}(Z|X_i) \tag{2.4}$$

for $\delta_i = 1$ is the Schoenfeld (1982) residual where $\hat\beta$ is usually obtained using the partial likelihood.

2.2 Measure of explained variation

The measure R^2

Let us first assume Z of dimension one. In (2.3) the expectation $\mathcal{E}_\beta(Z|X_i)$ is worked out with respect to an exponentially tilted distribution. The stronger the regression effects the greater the tilting, and the smaller we might expect, on average, the values $r_i^2(\beta)$ to be when compared with the residuals under the null model $\beta = 0$. Based on these residuals, a measure of explained variation can be defined (O'Quigley and Flandre, 1994), which is analogous to the coefficient of determination for the linear model. In the absence of censoring the quantity $\sum_{i=1}^n r_i^2(\hat\beta)/n$ is a residual sum of squares, and can be viewed as the average discrepancy between the observed covariate and its expected value under the model, whereas $\sum_{i=1}^n r_i^2(0)/n$ is a total sume of squares, and can be viewed as the average discrepancy without a model. Since the semiparametric model leaves inference depending only on the failure time rankings, and being able to predict failure rankings of all the failed subjects is equivalent to being able to predict at each failure time which subject is to fail, it is sensible to measure the discrepancy between the observed covariate at a given failure time and its expected value under the model. Thus we can define $\mathcal{I}(b)$ for $b = 0, \beta$ by

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty \{Z_i(t) - \mathcal{E}_b(Z|t)\}^2 dN_i(t) = \sum_{i=1}^{n} r_i^2(b). \tag{2.5}$$

We now define

$$R^2(\beta) = 1 - \frac{\sum_{i=1}^n r_i^2(\beta)}{\sum_{i=1}^n r_i^2(0)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)}, \tag{2.6}$$

This corresponds to the definition given by O'Quigley and Flandre (1994).

However, as pointed out in O'Quigley and Xu (2001), in order to completely eliminate any asymptotic dependence upon censoring, it is necessary to weight the squared Schoenfeld residuals by the increments of any consistent estimate of the marginal failure time distribution function F. Therefore, let $\hat{F}$ be the left-continuous Kaplan-Meier (KM) estimate of F, and define $W(t) = \hat{S}(t)/\sum_1^n Y_i(t)$ where $\hat{S} = 1 - \hat{F}$. Then $W(t)$ is a non-negative predictable stochastic process and, assuming there are no ties, it is straightforward to verify that $W(X_i) = \hat{F}(X_i+) - \hat{F}(X_i)$ at each observed failure time X_i, i.e. the jump of the KM curve. In practice, ties, if they exist, are split randomly. So, in place of the above definition of $\mathcal{I}(b)$ for $b = 0, \beta$, we use a more general definition in which

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty W(t)\{Z_i(t) - \mathcal{E}_b(Z|t)\}^2 dN_i(t) = \sum_{i=1}^{n} \delta_i W(X_i) r_i^2(b). \tag{2.7}$$

We now define

$$R^2(\beta) = 1 - \frac{\sum_{i=1}^n \delta_i W(X_i) r_i^2(\beta)}{\sum_{i=1}^n \delta_i W(X_i) r_i^2(0)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)}, \tag{2.8}$$

The definition given by O'Quigley and Flandre (1994) would be the same as above if we defined $W(t)$ to be constant and, of course, the two definitions coincide in the absence of censoring. The motivation for the introduction of the weight $W(t)$ is to obtain large sample properties of R^2 that are unaffected by an independent censoring mechanism. Viewing R^2 as a function of β turns out to be useful in theoretical studies. In practice, we are mostly interested in $R^2(\hat{\beta})$ where $\hat{\beta}$ is a consistent estimate of β such as the partial likelihood estimate.

In general when $Z(t)$ is a $p \times 1$ vector, the dependence of the survival time variable on the covariates is via the prognostic index (Andersen et al., 1983; Altman and Andersen, 1986)

$$\eta(t) = \beta' Z(t).$$

So we can imagine that each subject in the study is now labelled by η. R^2 as a measure of explained variation or, predictive capability, should evaluate how well the model predicts which individual or equivalently, its label, is chosen to fail at each observed failure time. This is equivalent to predicting the failure rankings given the prognostic indices. When $p = 1$, Z is equivalent to η, therefore we can construct the R^2 using residuals of the Z's. But for $p > 1$, the model does not distinguish between different vector Z's as long as the corresponding η's are the same. So instead of residuals of Z, we define the multiple coefficient using residuals of η. Consider then

$$R^2(\beta) = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)} \tag{2.9}$$

where

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty W(t)\{\eta_i(t) - \beta' \mathcal{E}_b(Z|t)\}^2 dN_i(t) = \sum_{i=1}^{n} \delta_i W(X_i)\{\beta' r_i(b)\}^2. \tag{2.10}$$

Population parameter Ω^2

The population parameter $\Omega^2(\beta)$ of $R^2(\hat{\beta})$ was originally given in O'Quigley & Flandre (1994). Let

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) e^{\beta' Z_i(t)} Z_i(t)^{\otimes r}, \quad s^{(r)}(\beta, t) = ES^{(r)}(\beta, t), \tag{2.11}$$

for $r = 0, 1, 2$. Here $a^{\otimes 2} = aa'$ and $a \otimes b = ab'$ for vectors a and b. We assume that conditions **A**-**C** of the appendix hold. Notice that $\mathcal{E}_\beta(Z|t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$. Let

$$J(\beta, b) = \int w(t)\beta' \left\{ \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - 2\frac{s^{(1)}(\beta, t) \otimes s^{(1)}(b, t)}{s^{(0)}(\beta, t) s^{(0)}(b, t)} + \frac{s^{(1)}(b, t)^{\otimes 2}}{s^{(0)}(b, t)^2} \right\} \beta s^{(0)}(\beta, t) \lambda_0(t) dt \tag{2.12}$$

where $w(t) = S(t)/s^{(0)}(0, t)$. Then

$$\Omega^2(\beta) = 1 - \frac{J(\beta, \beta)}{J(\beta, 0)}. \tag{2.13}$$

Notice that although (2.9) and (2.13) are not defined for $\beta = 0$, the limits exist and are equal to zero as $\beta \to 0$. So we can define $R^2(0) = \Omega^2(0) = 0$.

It has been shown (Xu, 1996) that $\Omega^2(\beta)$ is unaffected by an independent censorship mechanism, i.e. when C is independent of T and Z, and in this case it can be written (O'Quigley and Flandre, 1994)

$$\Omega^2(\beta) = 1 - \frac{\int E_\beta\{[Z(t) - E_\beta(Z(t)|t)]^2|t\}dF(t)}{\int E_\beta\{[Z(t) - E_0(Z(t)|t)]^2|t\}dF(t)}. \tag{2.14}$$

If, in addition, Z is time-invariant, we will see that $\Omega^2(\beta)$ has the interpretation of the proportion of explained variation (Section 3.2). O'Quigley and Flandre showed that, having standardized for the mean and the variance, $\Omega^2(\beta)$ depends only relatively weakly on different covariate distributions, and values of $\Omega^2(\beta)$ appear to give a good reflection of strength of association as measured by β and tend to 1 for high but plausible values of β (see also Table 1). Their numerical results support the conjecture that Ω^2 increases with the strength of effect, thereby agreeing with the third stipulation of Kendall (1975, p.4) for a measure of rank correlation. The conjecture was proven to be true in Xu (1996); see also Section 3.1. The first two stipulations were that perfect agreement or disagreement should reflect itself in a coefficient of absolute value 1; the third stipulation that for other cases the coefficient should have absolute value less than 1, and in some acceptable sense increasing values of the coefficient should correspond to increasing agreement between the ranks. Here we have a squared coefficient, and Kendall's stipulations are considered in a broader sense because we are not restricted to the ranks of the covariates in the semiparametric context.

3 Properties and interpretation

In this section we show that the measure defined above has the desired properties and the interpretation as a measure of explained variation. We omit all the proofs here. They can be found in Xu (1996).

3.1 Properties of R^2 and Ω^2

The R^2 defined above, can be shown to have the following properties:

(1) $R^2(0) = 0$;

(2) $R^2(\hat{\beta}) \leq 1$;

(3) $R^2(\hat{\beta})$ *is invariant under linear transformations of* Z *and monotonically increasing transformations of* T;

(4) $R^2(\hat{\beta})$ *consistently estimates* $\Omega^2(\beta)$. *In particular,* $\mathcal{I}(\hat{\beta})$ *and* $\mathcal{I}(0)$ *consistently estimate* $J(\beta, \beta)$ *and* $J(\beta, 0)$, *respectively.*

(5) $R^2(\hat{\beta})$ *is asymptotically normal.*

Note that in finite samples R^2, unlike Ω^2 below, cannot be guaranteed to be non-negative. A negative value for R^2 would correspond to the unusual case in which the best fitting model, in a least squares sense, provides a poorer fit than the null model. Our experience is that $R^2(\hat{\beta})$ will only be slightly negative in finite samples if $\hat{\beta}$ is very close to zero.

Similarly, we have the following properties for Ω^2:

(1) $\Omega^2(0) = 0$;

(2) $0 \leq \Omega^2(\beta) \leq 1$;

(3) $\Omega^2(\beta)$ *is invariant under linear transformations of* Z *and monotonically increasing transformations of* T;

(4) *for a scalar* β, $\Omega^2(\beta)$ *as a function of* β, *increases with* $|\beta|$; *and as* $|\beta| \to \infty$, $\Omega^2(\beta) \to 1$.

From the last property above, one can show that Ω^2 increases with the predictability of survival rankings, i.e. $P(T_i > T_j)$ for given Z_i and Z_j (assuming without loss of generality that $\beta > 0$). This corresponds to Kendall's third stipulation, in the context of semiparametric Cox regression.

3.2 Interpretation

In order to be completely assured before using R^2 in practice it is important to know that R^2 is consistent for Ω^2, that $\Omega^2(0) = R^2(0) = 0$, $\Omega^2(\infty) = 1$, that Ω^2 increases as strength of effect increases, and that Ω^2 is unaffected by an independent censoring mechanism. This enables us to state that an Ω^2 of 0.4 translates greater predictability than an Ω^2 of 0.3. We do, however, need one more thing. We would like to be able to say precisely just what a value such as 0.4 corresponds to. That is the purpose of this subsection.

A sum of squares decomposition

In definition (2.9) of $R^2(\beta)$, $\sum_{i=1}^n \delta_i W(X_i)\{\beta' r_i(\beta)\}^2$ can be considered as a residual sum of squares analogous to the linear regression case, while $\sum_{i=1}^n \delta_i W(X_i)\{\beta' r_i(0)\}^2$ is the total sum of squares. So define

$$\begin{aligned} SS_{\text{tot}} &= \sum_{i=1}^n \delta_i W(X_i)\{\hat\beta' r_i(0)\}^2, \\ SS_{\text{res}} &= \sum_{i=1}^n \delta_i W(X_i)\{\hat\beta' r_i(\hat\beta)\}^2, \\ SS_{\text{reg}} &= \sum_{i=1}^n \delta_i W(X_i)\{\hat\beta' \mathcal{E}_{\hat\beta}(Z|X_i) - \hat\beta' \mathcal{E}_0(Z|X_i)\}^2. \end{aligned}$$

It can be shown that an asymptotic decomposition holds of the total sum of squares into the residual sum of squares and the regression sum of squares, i.e.

$$SS_{\text{tot}} \overset{\text{asymp.}}{=} SS_{\text{res}} + SS_{\text{reg}}, \tag{3.1}$$

the difference between the two sides of the equation converging to zero in probability as $n \to \infty$. So R^2 is asymptotically equivalent to the ratio of the regression sum of squares to the total sum of squares.

Explained variation

For time-invariant covariates and independent censoring, the coefficient $\Omega^2(\beta)$ has a simple interpretation in terms of explained variation, i.e.

$$\Omega^2(\beta) \approx 1 - \frac{E\{\text{Var}(Z|T)\}}{\text{Var}(Z)} = \frac{\text{Var}\{E(Z|T)\}}{\text{Var}(Z)}. \tag{3.2}$$

Here we again omit the technical argument leading to the above approximation, but rather show the simulation results of Table 1. Indeed there is nothing to stop us defining explained variation as in the right hand side of (3.2), since the marginal distribution of Z and T can be estimated by the empirical and the KM estimator, while the conditional distribution of Z given $T = t$ by the $\{\pi_i(\hat\beta, t)\}_i$. However we can see no advantage to this and recommend that all calculations be done via the Schoenfeld residuals, evaluated at $\beta = \hat\beta$ and $\beta = 0$.

Table 1: Ω^2 as explained variation

covariate*	c	c	d	c	c	c	d
β	0	0.7	0.7	1.4	2.8	4.2	4.2
$R^2(\beta)$	0.0002	0.0990	0.0979	0.2844	0.5887	0.7577	0.8728
$\text{Var}\{E(Z\|T)\}/\text{Var}(Z)$	0.0018	0.0998	0.0985	0.2848	0.5889	0.7578	0.8728

* Covariate distribution: d – binary, c – uniform. Data are simulated under the same mechanism as in Section 4.

4 Simulation results

In this section through simulations we compare the behavior of R^2 with some of the measures mentioned in the introduction. We make use of some of the results from table II of Schemper and Stare (1996). In table 2, data are generated with hazard function $\lambda(t) = \exp(-\beta' Z)$, where $\beta = 0, \log 2, \log 4, \log 16, \log 64$, and Z distributed as either uniform $[0, \sqrt{3}]$ ('c') or dichotomous 0,1 with equal probabilities ('d'). These two covariate distributions have identical variances and thus allow comparison of the results for continuous and dichotomous covariates. Censoring mechanisms are uniform $[0, \tau]$, where τ is chosen to achieve a certain percentage of censoring.

In the table, R^2 is the measure of O'Quigley and Xu (2001), and R^2_{OQF} the original proposal from O'Quigley and Flandre (1994). The measures based on information gain are not listed here; they are studied in a separate manuscript. We simply mention that they have close agreement with the R^2 measures considered here. The next five columns are from Schemper and Stare (1996), where the measure V_2 is from Schemper (1990, 1994), r^2_{pr} is from Schemper and Kaider (1997), R^2_M is from Maddala (1983), and D^2 and 'KS' are from Korn and Simon (1990) with 'KS' based on quadratic loss.

From table 2 we see that R^2 and R^2_{OQF} are very close. The same is true in our experience with smaller sample sizes. This suggests that in practice we may able to use either, although theoretically only R^2 is consistent for Ω^2 in the presence of censoring. The measures V_2 and R^2_M are clearly affected by censoring, and would give uninterpretable results if used with censored survival data. The censoring also has an impact upon the KS measure. Unlike all the other measures included, this measure does not remain invariant to monotone increasing

Table 2: A simulated comparison of different measures ($n = 5000$)

$\exp(\beta)$	% censored	covariate	R^2	R^2_{OQF}	V_2	r^2_{pr}	R^2_M	D^2	KS
1	0%	c	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	50%	c	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	90%	c	0.002	0.002	0.000	0.000	0.000	0.000	0.000
2	0%	c	0.098	0.098	0.094	0.092	0.113	0.042	0.101
	50%	c	0.101	0.105	0.059	0.093	0.063	0.044	0.088
	90%	c	0.104	0.104	0.009	0.074	0.010	0.040	0.015
	0%	d	0.099	0.099	0.097	0.096	0.112	0.032	0.095
	50%	d	0.105	0.108	0.061	0.096	0.064	0.035	0.089
	90%	d	0.112	0.108	0.010	0.076	0.011	0.031	0.016
4	0%	c	0.281	0.281	0.276	0.272	0.308	0.132	0.231
	50%	c	0.303	0.316	0.187	0.274	0.198	0.142	0.267
	90%	c	0.325	0.334	0.034	0.278	0.042	0.149	0.063
16	0%	c	0.586	0.586	0.591	0.584	0.607	0.323	0.354
	50%	c	0.623	0.644	0.463	0.584	0.461	0.370	0.564
	90%	c	0.703	0.708	0.107	0.585	0.115	0.376	0.188
64	0%	c	0.757	0.757	0.762	0.754	0.762	0.463	0.397
	50%	c	0.790	0.806	0.644	0.730	0.626	0.537	0.717
	90%	c	0.863	0.864	0.195	0.694	0.180	0.536	0.321
	0%	d	0.870	0.870	0.723	0.707	0.685	0.236	0.319
	50%	d	0.873	0.885	0.539	0.718	0.638	0.370	0.861
	90%	d	0.941	0.937	0.070	0.701	0.118	0.257	0.135

transformation of time; it is most useful when the time variable provides more information than just an ordering. The values of r_{pr}^2 tend to be slightly lower than R^2 and R^2_{OQF}, although the strength of association reflected is similar. The measure r_{pr}^2 requires more computation than all the other ones in the table because of the multiple imputation technique employed. In addition, like some of the other proposed measures, statistical properties of the measure have yet to be studied. Finally the measure D^2 gives rather small values, even when the hazard ratio is as high as 64. The values of the measure are further reduced if the covariate distribution is discrete.

5 Extensions

5.1 Partial coefficients

The partial coefficient can be defined via a ratio of multiple coefficients of different orders. Specifically, and in an obvious change of notation just for the purposes of this subsection, let $R^2(Z_1, \ldots, Z_p)$ and $R^2(Z_1, \ldots, Z_q)$ $(q < p)$ denote the multiple coefficients with covariates Z_1 to Z_p and covariates Z_1 to Z_q, respectively. Note that $R^2(Z_1, \ldots, Z_p)$ is calculated using $\hat{\beta}_1, \ldots, \hat{\beta}_p$ estimated when $Z_1, \ldots, Z_p$ are included in the model, and $R^2(Z_1, \ldots, Z_q)$ using $\hat{\beta}_{10}, \ldots, \hat{\beta}_{q0}$ estimated when only $Z_1, \ldots, Z_q$ are included. Define the partial coefficient $R^2(Z_{q+1}, \ldots, Z_p | Z_1, \ldots, Z_q)$, the correlation after having accounted for the effects of Z_1 to Z_q by

$$1 - R^2(Z_1, \ldots, Z_p) = [1 - R^2(Z_1, \ldots, Z_q)][1 - R^2(Z_{q+1}, \ldots, Z_p | Z_1, \ldots, Z_q)]. \tag{5.1}$$

The above coefficient, motivated by an analogous expression for the multivariate normal model, makes intuitive sense in that the value of the partial coefficient increases as the difference between the multiple coefficients increases, and takes the value zero should this difference be zero. Partial Ω^2 can be defined in a similar way.

We can also derive definition (5.1) directly. Following the discussion of multiple coefficients, we can use the prognostic indices obtained under the model with $Z_1, \ldots, Z_p$ and that with $Z_1, \ldots, Z_q$. This would be equivalent to defining $1 - R^2(Z_{q+1}, \ldots, Z_p | Z_1, \ldots, Z_q)$ as $\mathcal{I}(Z_1, \ldots, Z_p)/\mathcal{I}(Z_1, \ldots, Z_q)$, the ratio of the numerators of $1 - R^2(Z_1, \ldots, Z_p)$ and $1 - R^2(Z_1, \ldots, Z_q)$. However, since the two numerators are on different scales, being inner products of vectors of different dimemsions, their numerical value require standardization. One natural way to standardize is to divide these numerators by the denominators of $1 - R^2(Z_1, \ldots, Z_p)$ and $1 - R^2(Z_1, \ldots, Z_q)$, respectively. This gives definition (5.1).

Partial coefficients in O'Quigley and Flandre (1994) were defined using a single component of the covariate vector instead of the prognostic index. Although our limited data experience did not show any important discrepancies between that definition and (5.1), there seems to be some arbitrariness as to which component of the vector to use. Furthermore the prognostic index should reflect the best prediction a given model can achieve in the sense we described before. Our recommendation is to use (5.1) as the partial coefficient.

5.2 Stratified model

The partial coefficients of the previous section enable us to assess the impact of one or more covariates while adjusting for the effects of others. This is carried out in the context of the assumed model. It may sometimes be preferable to make weaker assumptions than the full model and adjust for the effects of other multilevel covariates by stratification. Indeed it can be interesting and informative to compare adjusted R^2 measures, the adjustments having been made either via the model or via stratification. For the stratified model the definitions of Section 2 follow through readily. To be precise, we define a stratum specific residual for stratum s $(s = 1, ..., S)$, where, in the following, a subscript is in place of i means the i th subject in stratum s. Thus we have

$$r_i(b; s) = Z_{is}(X_{is}) - \mathcal{E}_b(Z|X_{is}) \tag{5.2}$$

where $\mathcal{E}_b(Z|X_{is})$ is averaged within stratum s over the risk set at time X_{is}, and we write

$$\mathcal{I}(b) = \sum_i \sum_s \int_0^\infty W(t)\{Z_{is}(t) - \mathcal{E}_b(Z|t)\}^2 dN_{is}(t) = \sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(b, s). \tag{5.3}$$

From this we can define

$$R^2(\beta) = 1 - \frac{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(\beta, s)}{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(0, s)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)} \tag{5.4}$$

Note that we do not use a stratum specific $W(t)$ and, as before, we work with an assumption of a common underlying marginal survival distribution. The validity of this hinges upon an independent, rather than a conditionally independent, censoring mechanism. Under a conditionally independent censoring mechanism, a weighted Kaplan-Meier estimate (Murray and Tsiatis, 1996) of the marginal survival distribution may be used instead.

5.3 Other relative risk models

It is straightforward to generalize the R^2 measure to other relative risk models, with the relative risk of forms such as $1 + \beta z$ or $\exp\{\beta(t)z\}$. Denote $r(t; z)$ a general form of the relative risk. Assume that the regression parameters involved have been estimated, and define $\pi_i(t) = Y_i(t)\hat{r}(t; Z_i)/\sum_{j=1}^n Y_j(t)\hat{r}(t; Z_j)$. Then we can similarly define $\mathcal{E}_\beta(Z|t)$ and form the residuals, thereby defining an R^2 measure similar to (2.8). In addition, it can be shown that under an independent censorship, the conditional distribution of $Z(t)$ given $T = t$ is consistently estimated by $\{\pi_i(t)\}_i$, so properties such as being unaffected by an independent censorship are maintained.

It is particularly interesting to study the use of such an R^2 measure under the time-varying regression effects model, where the relative risk is $\exp\{\beta(t)z\}$. Different approaches have been proposed to estimate $\beta(t)$ (Sleeper and Harrington, 1990; Zucker and Karr, 1990; Murphy and Sen, 1991; Gray, 1992; Hastie and Tibshirani, 1993; Verweij and Van Houwelingen, 1995; Sargent, 1997; Gustafson, 1998; Xu and Adak, 2001). In this case we can use R^2 to compare the predictability of different covariates as we do under the proportional hazards model; we can also use it to guide the choice of the amount of smoothness, or the "effective degrees of freedom" as it is called by the some of the aforementioned authors, in estimating $\beta(t)$. As a brief illustration, suppose that we estimate $\beta(t)$ as a step function, and that we are to choose between two different partitions of the time axis, perhaps one finer than the other. Denote the two estimates obtained under these two partitions by $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$, the latter corresponding to the finer partition. We can measure the extra amount of variation explained by fitting $\hat{\beta}_2(t)$ versus fitting $\hat{\beta}_1(t)$, by

$$R^2_{\text{ex}} = 1 - \frac{\mathcal{I}(\hat{\beta}_2(\cdot))}{\mathcal{I}(\hat{\beta}_1(\cdot))}. \tag{5.5}$$

This can be thought of as a partial coefficient, if we look at the "dimension" of $\beta(t)$ through time. The use of R^2_{ex} in estimating $\beta(t)$ was recently adopted in Xu and Adak (2001).

6 Application to cancer studies

In the following, through practical examples in cancer research, we illustrate how the R^2 measure can throw light on the complex interrelations between the risk factors and prognosis. Here we emphasize the use of R^2, while omitting the detailed data analysis with regard to other aspects of model diagnostics.

Example 1 The first example concerns the breast cancer study described briefly in the introduction. We first fit different proportional hazard models with a single covariate and calculated the R^2 as defined in (2.9). These results are summarized in table 3. All variables are highly significant. The predictive power though, is quite different. Stage and tumor size, as one might expect, have reasonably high predictability. Histology grade also has predictive power, although this covariate has been shown to have a nonproportional regression effect. We investigated a more complex model in which the coefficient for histology was allowed to decay with time. The value of R^2 increased from 0.116 to 0.24, the improvement in explained variation reflecting an improvement in fit. This case also underlines the relationship between predictability and goodness-of-fit. On the other hand, age has very weak predictive capability, though significant. This estimated weak effect could be due to: 1) a population weak effect, or 2) a suboptimal coding of the covariate. We investigated this second possibility via two recoded models. The first, making a strong trend assumption, coded age as 1 (0-33), 2 (34-40) and 3 (41 and above). The second model, making no assumptions about trend, used two binary variables to code the three groups. All three models gave very similar values of R^2. In consequence only the simplest model is retained for subsequent analysis, i.e. the age groups 1-3. In addition, we calculated the multiple R^2 for a set of nested models. These results are illustrated in table 4. Table 4 also contains the values of the partial R^2 defined in (5.1), when each additonal covariate is added to the existing model. The partial coefficient for tumor size having accounted for the other four variables is 0.006, suggesting that the extra amount of variation in survival explained by the patient's tumor size is quite limited. The use of partial R^2 is further explored in Example 3.

Example 2 In a study on prognostic factors in gastric cancer (Rashid et al., 1982) certain acute phase reactant proteins were measured pre-operatively. Five covariates were studied: stage together with the proteins α_1-anti chimotrypsin (ACT), carcino embryonic antigen

Table 3: Breast cancer – univariate analysis

covariate	$\hat{\beta}$	p-value	R^2
age	-0.24	<0.01	0.005
hist	0.37	<0.01	0.12
stage	0.53	<0.01	0.20
prog	-0.73	<0.01	0.07
size	0.02	<0.01	0.18

Table 4: Breast cancer – multivariate analysis

covariates	R^2	partial R^2
age	0.01	
age and hist	0.12	0.12
age, hist, and stage	0.26	0.16
age, hist, stage, and prog	0.33	0.09
age, hist, stage, prog, and size	0.33	0.01

(CEA), C-reactive protein (CRP) and α_1 glyco-protein (AGP). Surgery is needed in order to determine the stage of the cancer, a clinical factor known to strongly influence survival, and one of the purposes of the study was to find out how well the four protein covariates, available pre-operatively, are able to explain survival in the absence of information on stage. A logarithm transformation for CEA was found to be necessary. This is also reflected in a R^2 increasing from 0.10 to 0.20 after the transformation. Table 5 shows that each of the five covariates has reasonable predicting power, with R^2 for stage alone to be 0.48. A direct calculation of sample correlation shows that ACT, CRP and AGP are highly correlated, which is supported by biological evidence. In addition, fitting the Cox model with all four protein covariates shows that CRP and AGP are no longer significant in the presence of the other covariates. These two variables were dropped from further study. The value of R^2 for a model with ACT and log(CEA) is 0.37; this increases to 0.54 when stage is also included, and the corresponding partial R^2 is equal to 0.27. In conclusion, there is strong prognostic information in the pre-operative measurements ACT and log(CEA), but this only partially captures the information contained in stage.

Example 3 Our third example was motivated by the increasing number of correlative stud-

Table 5: Gastric cancer – univariate analysis

covariate	$\hat{\beta}$	p-value	R^2
stage	1.78	<0.01	0.48
ACT	2.26	<0.01	0.29
log(CEA)	0.30	<0.01	0.20
CRP	0.02	<0.01	0.26
AGP	0.70	<0.01	0.14

ies carried out in cancer research to relate the outcome with multi-dimensional molecular and genetic markers. Our data come from a clinical trial (EST 9486) of multiple myeloma conducted by the Eastern Coorperative Oncology Group (Oken et al, 1999). The trial collected laboratory measurements on patients' myeloma cells, including measurements from the blood or serum (albumin, β_2 microglobin, creatinine, immunoglobins IgA and IgG, percent plasma cells, and hemoglobin); characteristics of the circulating myeloma cells (plasma cell labelling index, IL-6 receptor status, and C-reative protein); and kappa light chain. The study assigned subjects to three randomized treatment arms, but no significant survival difference was found across the three arms. Here we include a randomly selected group of 295 patients, on whom a particular chromosomal abnormality, the possible deletion of the short arm of chromosome 13 (denoted by 13q-), was measured by flouresecent in-situ hybridization (FISH) in the laboratory of R. Fonseca at the Mayo Clinic. We also include the traditional Durie-Salmon stage which was routinely used to predict prognosis in multiple myeloma before the availability of assays to measure genetic and other molecular abnormalitiies of the myeloma cells.

Univariate Cox regression analysis indicates that all of the above 13 covariates are associated with patients' survival times (p-value < 0.23), and most of them are highly significant. Table 6 shows the estimated regression effects and the standard errors, and the univariate R^2 coefficients. As we see the predictability by an individual marker is generally low, with the highest R^2 of 0.08 from plasma cell labelling (PCL) index. However, when all the covariates are included in a multiple Cox model, only six of them remain significant (p-value < 0.08), with the multiple $R^2 = 0.20$. In particular, the traditional staging system is no longer significantly predictive of survival given the laboratory measurements. Leaving out

Table 6: Myeloma – univariate analysis

covariate	$\hat{\beta}$	se($\hat{\beta}$)	R^2
creatinine	0.66	0.16	0.05
plasma	0.43	0.12	0.04
IL-6	0.35	0.14	0.02
C-reactive	0.53	0.17	0.02
a13q	0.22	0.12	0.01
hemoglobin	-0.30	0.13	0.03
albumin	-0.39	0.14	0.03
IgG	-0.15	0.12	0.01
IgA	0.16	0.14	0.01
kappa	-0.26	0.12	0.01
stage	-0.18	0.12	0.004
β_2 microglobin	0.48	0.13	0.03
PCL index	0.59	0.13	0.08

the non-significant variables in a Cox model gives $R^2 = 0.18$. As an illustration of variable selection using R^2, we build hierarchical models starting with PCL index which has the highest univariate R^2. We then choose the variable among the rest five that has the highest partial R^2, and so on. Table 7 gives the nested models and the corresponding R^2's.

The above examples serve the purposes of illustration and help underline our arguments that R^2 is useful. They are not in any sense thorough, and, in practical applications, deeper study would be useful. For instance, the combinatorial problem of examining all possible subgroups of different sizes raises both statistical and computational challenges. The statistical question, also present in the limited analyses above, is that of bias, or inflation, of multiple R^2 away from zero, when viewed as an estimate of the corresponding multiple Ω^2. This question is not specific to the survival setting and arises in the standard case of linear regression. Bias reduction techniques used there, such as bootstrap resampling, would also be helpful for our application. In the myeloma example it is quite likely that the small observed increases in R^2 when adding further variables to the vector (PLC, creatinine, plasma, a13q) are artifacts of the data rather than indications that similar increases also hold for Ω^2.

Table 7: Myeloma – nested models

covariates	R^2
PLC	0.08
PLC, creat	0.11
PLC, creat, plasma	0.13
PLC, creat, plasma, a13q	0.16
PLC, creat, plasma, a13q, β_2 mcrglb	0.17
PLC, creat, plasma, a13q, β_2 mcrglb, IL-6	0.18
all 13 variables	0.20

7 Conclusion

The R^2 measure studied in this paper, is a natural analogue to the usual R^2 for linear regression. It has the properties desirable for such a measure, as well as concrete interpretations including predictability of survival rankings, sums of squares decomposition and proportion of explained variation. The measure naturally accommodates time-dependent covariates, and can be easily computed after the proportional hazards regression model has been fitted. All that is required is the squaring and summing of the Schoenfeld residuals, under the null and the fitted models. Extensions to other relative risk models are straightforward. We recommend the measure for routine use.

Acknowledgement

The authors would like to thank Dr. Bernard Asselain of the Institut Curie, Paris, France for permission to analyse the breast cancer data. EST 9486 was supported by the ECOG grant CA23318.

APPENDIX

For showing large sample properties, we make the following assumptions which are similar to those in Andersen and Gill (1982):

A. (Finite interval). $\int_0^1 \lambda_0(t)dt < \infty$.

B. (Asymptotic stability). There exist a neighbourhood $\mathcal{B}$ of β such that 0 and β_0 belong to the interior of $\mathcal{B}$, and

$$\sup_{t\in[0,1]} \|nW(t) - w(t)\| \xrightarrow{\text{P}} 0,$$

$$\sup_{t\in[0,1],\beta\in\mathcal{B}} \|S^{(r)}(\beta,t) - s^{(r)}(\beta,t)\| \xrightarrow{\text{P}} 0,$$

for $r = 0, 1, 2, 3, 4$, where the arrows indicate convergence in probability with rate $n^{-1/2}$.

C. (Asymptotic regularity conditions). All functons in **B** are uniformly continous in $t \in [0,1]$; $s^{(r)}(\beta,t), r = 0, 1, 2, 3, 4$, are continuous functions of $\beta \in \mathcal{B}$, and are bounded on $\mathcal{B} \times [0,1]$; $s^{(0)}(\beta,t)$ is bounded away from zero.

REFERENCES

Altman, D.G. and Andersen P.K. (1986) A note on the uncertainty of a survival probability estimated from Cox's regression model. *Biometrika* **73**, 722-724.

Andersen, P.K., Christensen, E., Fauerholdt, L. and Schlichting, P. (1983) Measuring prognosis using the proportional hazards model. *Scand. J. Statist.* **10**, 49-52.

Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

Cox, D.R. (1972) Regression models and life tables (with discussion). *J.R. Statist. Soc. B* **34**, 187-220.

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

Freireich, E.O. et al. (1963) The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood* **21**, 699-716.

Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942-951.

Gustafson, P. (1998) Flexible Bayesian modelling for survival data. *Lifetime Data Analysis* **4**, 281-299.

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757-796.

Kendall, M. (1975) *Rank correlation methods.* 4th Edition, London and High Wycombe, Griffin.

Kent, J.T. (1983) Information gain and a general measure of correlation. *Biometrika* **70**, No.1, 163-174.

Kent, J.T. and O'Quigley, J. (1988) Measure of dependence for censored survival data. *Biometrika* **75**, 525-534.

Korn, E.L. and Simon,R. (1990) Measures of explained variation for survival data. *Statistics in Medicine* **9**, 487-503.

Maddala, G.S. (1983) *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, U.K.

Murray, S. and Tsiatis A.A. (1996) Nonparametric survival estimation using prognostic longitudianl covariates. *Biometrics* **52**, 137-151.

Murphy, S.A. and Sen, P.K. (1991) Time-dependent coefficients in a Cox-type regression model. *Stoch. Process. Applic.* **39**, 153-80.

Oken, M.M., Leong, T., Lenhard, R.E., Greipp, P.R., Kay, N.E., Van Ness, B., Keimowitz, R.M. and Kyle, R.A. (1999) The addition of interferon or high dose cyclophosphamide to standard chemotherapy in the treatment of patients with multiple myeloma. *Cancer* **86**, 957-968.

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

O'Quigley J., Flandre P. and Reiner E. (1999) Large sample theory for Schemper's measures of explained variation in the Cox model. *J. Roy. Statist. Soc (D)* **48**, 53-62.

O'Quigley, J. and Xu, R. (2001) Explained variation in proportional hazards regression. *Handbook of Statistics in Clinical Oncology* (ed. J. Crowley), 397-409. Marcel Dekker, Inc., New York.

Rashid, S.A., O'Quigley, J., Axon, A.T.R. and Cooper, E.H. (1982) Plasma protein profiles and prognosis in gastric cancer. *British Journal of Cancer* **45**, 390-394.

Sargent, D.J. (1997) A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis* **3**, 13-25.

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

Schemper, M. (1994) Correction: 'The explained variation in proportional hazards regression'. *Biometrika* **81**, 631.

Schemper, M. and Kaider, A. (1997) A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards. *Computational Statistics and Data Analysis* **23**, 467-476.

Schemper, M. and Stare, J. (1996) Explained variation in survival analysis. *Statistics in Medicine* **15**, 1999-2012.

Schoenfeld, D.A. (1982) Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239-41.

Sleeper, L.A. and Harrington, D.P. (1990) Regression splines in the Cox model with application to covariate effects in liver disease. *JASA* **85**, 941-949.

Verweij, J.A. and Van Houwelingen, H.A. (1995) Time-dependent effects of fixed covariates in Cox regression. *Biometrics* **51**, 1550-1556.

Xu, R. (1996) *Inference for the proportional hazards model.* Ph.D thesis of University of California, San Diego.

Xu, R. and Adak, S. (2001) Survival analysis with time-varying relative risks: a tree-based approach. *Methods of Information in Medicine* **40**, 141-147.

Xu, R. and O'Quigley, J. (1999) A R^2 type of measure of dependence for proportional hazards models. *Nonparametric Statistics* **12**, 83-107.

Zucker, D.M. and Karr, A.F. (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics* **18**, 329-353.