Chapter 13

Explained variation

13.1 Summary

Some suggestions on possible measures of explained variation which have appeared in the literature are considered. Following this an outline of the recommended approach is given. Leaning upon the theory of explained variation detailed in Chapter 2 and in particular 3.9 we show how a solid theory of explained variation for proportional and non-proportional hazards regression can be established. This contrasts with a substantial body of literature on this topic, almost entirely constructed around intuitive improvisations and ad-hoc modifications to sample based quantities gleaned from classical linear regression. The main reference here is the paper by O'Quigley and Flandre (1994) which showed how the Schoenfeld residuals provide the required ingredients for the task in hand. The properties of population quantities and sample based estimates have been studied thoroughly (O'Quigley and Xu 2001) and these provide the user with the necessary confidence for their practical use.

13.2 Motivation

Referring back to Chapter 2 and Section 3.9 it is clear that the concept explained variation is a fundamental one, directly quantifying the notion of predictive ability. This quantification is a consequence of the Chebyshev inequality. As an example of a practical setting in which we are motivated to look at this, consider 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 and the relative importance of the different prognostic factors. 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 our situation, in which inference is rank invariant with respect to monotonic transformations on time, then from Section 3.9, we can see that this implies evaluation of the explained variation in the covariate given time rather than, the apparently more natural, explained variation of time given the covariate. For normal models the two are the same anyway and, here, we would anticipate them as being very close. In addition, we have all that is needed if we prefer to consider the explained variation of time given the covariates.

It helps to keep in mind the implication of working with the conditional distribution of the covariate given time rather than the other way around. It means that explained variation, translated as predictability as a consequence of Chebyshev's inequality, refers to the predictability of the failure ranks. Absence of effect should then translate as 0% predictability; perfect prediction of the correct ordering of the survival ranks should translate as 100%; and intermediate values are to be interpretable as providing an ordered scale, any point of which indicates precisely the amount of predictive strength in the model. These concepts are outlined below.

13.3 Finding a suitable measure of R^2

Some suggestions in the literature

The R^2 measure of explained variability, or predictive capability, is well known under a normal linear model. As pointed out by Korn and Simon (1990), and in contrast to what is oftentimes taught and written, such measures are only indirectly concerned with fit. They are directly concerned with predictability. For the proportional hazards model some correlation measures were first suggested by Harrell (1986) although it turned out that his measures depend heavily on independent censoring and can not be easily interpreted. 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 in the observed survival times explained by the covariates.

The principal 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 accommodate 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 accommodates time-dependent covariates, and is computable using standard softwares for fitting the Cox model. We consider this measure in the following chapter.

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. In this context these disadvantages are quite severe. Schemper (1990, 1994) introduced the concept of individual survival curves for each subject, with the model and without the model. Interpretation is very difficult. As with the Harrell measure, the Schemper measures depend on censoring, even when the censoring mechanism is completely independent of the failure mechanism. Schemper and Kaider (1997) proposed to estimate the correlation coefficient between failure rankings and the covariates via multiply imputing the censored failure times. Although numerically complex, and, again, not readily affording any clear interpretation, this latter coefficient of Schemper and Kaider shows promise and may be worthy of

further study. It is possible to remove the dependence on the censoring and this has been considered by O'Quigley, Flandre and Reiner (1999) and Schemper and Henderson (2000).

Distance measures

Explained variation is clearly based on a measure of distance. Some authors have preferred to directly address the question of predictive ability of any model via classes of distance measures. This is the case for Harrell (1986), Korn and Simon (1990), Schemper (1990, 1992) and Graf and Schumacher (1995). Apart from the measure of Harrell, which relates to measures of information gain described in the following chapter, all of these measures relate to those described by Schemper.

In this description of the Schemper measures we keep to his notation (Schemper 1990) in order to facilitate any comparative study the reader may be interested in carrying out. Schemper defined S_{ij}, interpretable as an "empirical survivorship function" per individual, for subject i at observed failure time point t_j ($j = 1, \ldots, k_i$). The quantity k_i will be the total number of failures should individual i correspond to a failure; otherwise k_i is the number of failures occurring prior to the censoring time of the individual i. $S_{ij} = 1$ for individual i at all time points t_j for which the individual is still alive, drops to 0.5 at the point at which the individual fails, and thereafter $S_{ij} = 0$. Note that changing the definition of S_{ij} so that it drops to zero rather than 0.5 at the observed failure time will have a negligible impact in practice and an impact approaching zero as sample size (number of failures) increases.

Denote further $\bar{S}_j$ to be the Kaplan-Meier estimate of survival at time t_j and $\bar{S}_{ij}$ the estimate of survival for individual i at time point t_j derived from the proportional hazards model. Two different measures of the *proportion of variability explained* were suggested, V_1 and V_2 where, for $\ell = 1, 2$:

Definition 13.1 *Schemper's proportion of variability explained is*

$$V_\ell = 1 - \frac{\sum k_i^{-1} \sum |S_{ij} - \bar{S}_{ij}|^\ell}{\sum k_i^{-1} \sum |S_{ij} - \bar{S}_j|^\ell}; \quad \ell = 1, 2. \tag{13.1}$$

For an exponential model and different relative risks, values of V_1 and V_2 were tabulated on the basis of a single large simulation (Schemper 1990). The entries for V_2 turned out not to be based on a sum of

squares, as the above expression and Schemper's original paper indicate, but in fact on a rather less classical squared sum (Schemper 1994). Thus, the original definition for V_2 was considered to be in error by Schemper (1994) and replaced by an alternative one, say V_2^*, for when $\ell = 2$, replacing $\sum |S_{ij} - \bar{S}_{ij}|^\ell$ by $\sum (k_i^{-1} \sum |S_{ij} - \bar{S}_{ij}|)^2$ in the numerator and $\sum k_i^{-1} \sum |S_{ij} - \bar{S}_j|^\ell$ by $\sum (k_i^{-1} \sum |S_{ij} - \bar{S}_j|)^2$ in the denominator. There is something unusual, requiring further justification it would seem, in working with distances defined in terms of squared sums rather than sums of squares. The merits of such a definition were not detailed by Schemper (1994) although subsequent work (O'Quigley, Flandre and Reiner 1999; Schemper and Henderson 2000) suggest the original definition should be retained as the correct one. In support of this is the interesting observation that, for an exponential model and no censoring, the population equivalents of V_1 and V_2 converge to the same quantity.

Schemper's coefficients can be seen to depend on the unknown independent censoring mechanism (O'Quigley, Flandre and Reiner 1999, Schemper and Henderson 2000). This can however be remedied and we look at this in a later section. The Schemper coefficients are generally bounded by a number strictly less than one. This is also true in the uncensored case and, for the cases studied by Schemper (1990), the population values of V_1 and V_2 are bounded by 0.5.

Relationship between distance measures

Discussion of the relationships between different coefficients based on some measure of distance is given in Graf and Schumacher (1995). A study of the Schemper proposal and its large sample properties is enough to deduce the properties we would anticipate from closely associated measures. We return to this in Section 13.10 and point out here the way in which these coefficients are connected. It is useful to consider the population equivalents of V_1 and V_2 and we do this by considering the probability limits of the numerator and denominator in definition 13.1. If, for $\ell = 1, 2$, the numerator converges in probability to N_ℓ and the denominator to D_ℓ then we can study the population parameter θ_ℓ where $\theta_\ell = 1 - D_\ell^{-1} N_\ell$. We look at this in more detail in Section 13.10. For now we simply consider the form of N_ℓ as this brings out the relationship between the distance measures.

Korn and Simon (1990) considered squared error to be a particular kind of loss function and therefore other kinds of loss function, such

as absolute error, might also be considered. The main development is around integrated squared error loss. For the numerator in their expression, let's call it N_{KS} here, we have

$$\tilde{N}_{KS} = \int\int \tilde{S}(u|z)\{1-\tilde{S}(u|z)\}dudH_n(z)\,. \qquad (13.2)$$

In the absence of censoring, for the population equivalent of V_2, we can construct a theoretical numerator, $\tilde{N}_2$ given by

$$\tilde{N}_2 = \int\int\int \{Y_t(u)-\tilde{S}(u|z)\}^2 d\tilde{F}(u)d\tilde{F}(t|z)dH_n(z).$$

In the uncensored case then the distance measures are closely related. The differences arise as a result of the weightings. For the Schemper coefficients these are given in terms of increments in $\tilde{F}(t)$ rather than increments in t itself. This we deduce from taking the above integral one step further where we see that:

$$\tilde{N}_2 = \int\int \tilde{S}(u|z)\{1-\tilde{S}(u|z)\}d\tilde{F}(u)dH_n(z)\,, \qquad (13.3)$$

which we can then compare with Equation 13.2. The same conclusion has also been obtained by Graf and Schumacher (1995). Note that monotonic transformations of t would typically impact the Korn and Simon measures, whereas the increments in $\tilde{F}(t)$, and thereby V_ℓ itself, remain unaffected. Given that inference under the proportional hazards model has this invariance property, it may be considered a desirable property of V_ℓ. Furthermore, for the broad class proposed by Korn and Simon (1990), it would be straightforward to extend their measures by adopting such a modification, in order to accommodate such a property if deemed necessary.

Recommended approach

The most transparent approach, interpretable in terms of explained variation, is that described by O'Quigley and Flandre (1994). This approach, in tune with the general theory of Section 3.9, studies the explained variation in T given the covariate vector Z, or, in order to maintain rank invariance, the explained variation of the prognostic index (Z alone in the univariate case) given T. If we stray from this we lose interpretability and, although many of the other suggestions have merit, they can run into all sorts of problems such as unknown

bounds on the index, negative values, strong dependence on the censoring, even when independent of the failure mechanism and, simply, no way to interpret them. Thus, a value of 0.03, under one set of circumstances, may indicate a stronger effect than a value of 0.5, obtained under a different set of circumstances. A more solid approach can be constructed by keeping the basic theory in mind from Section 3.9. Leaning on that basic theory we can anticipate obtaining indices with meaningful properties. Even so, it is still important to investigate any properties deemed desirable, and not automatically inherited by virtue of Section 3.9.

Our recommended approach is essentially that outlined in O'Quigley and Flandre (1994). Their motivation came from linear regression where we denote $r_i(\hat{\beta})$ to be the fitted residual, i.e., the difference between the observation and its model based expectation evaluated under $\beta = \hat{\beta}$. The null residual $r_i(0)$ obtains by putting instead $\beta = 0$ and this corresponds to replacing all expectations by the overall mean. Next we calculate the average squared deviation of the observations from their model based predictions, $\sum r_i^2(\beta)/n$, leading to the well known expression for R^2, written as $R^2(\beta)$ in order to make explicit the dependence on $\hat{\beta}$, from

$$R^2(\beta) = 1 - \frac{\sum r_i^2(\beta)}{\sum r_i^2(0)}. \tag{13.4}$$

Some additional work was needed in order for the R^2 measure of O'Quigley and Flandre to be consistent in general situations. This is achieved by weighting things correctly and this is described below. We discuss all the needed statistical properties for the measure including obtaining confidence intervals with coverage properties asymptotically the same as those for the regression coefficient estimate itself. A sum of squares decomposition, an expression for explained variation and the relationship between increasing values of the measure and predictability of the survival ranks all help form the basis for a more solid interpretation. Via simulations we compare this measure with some of the measures mentioned above. Those aspects particular to the multicovariate case are examined more closely and some general recommendations are given. The measure can also be easily extended to other relative risk models.

13.4 An R^2 measure based on Schoenfeld residuals

Recall the Schoenfeld residuals as the discrepancy between the observed value of the covariate, viewed of as having been sampled at time point X_i and its expected value,

$$r_i(\beta) = Z_i(X_i) - \mathcal{E}_\beta(Z|X_i), \tag{13.5}$$

for $\delta_i = 1$ at each observed failure time X_i. 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, analogous to the coefficient of determination for the linear model, can be defined (O'Quigley and Flandre 1994).

Since the semiparametric model leaves inference invariant under monotonic increasing transformations of the time axis, and being able to predict at each failure time which subject is to fail is equivalent to being able to predict failure rankings of all the failed subjects, it is sensible to measure the discrepancy between the observed covariate at a given failure time and its expected value under the model. In the absence of censoring the quantity $\sum_{i=1}^n r_i^2(\hat{\beta})/n$ can be viewed as the average discrepancy between the observed covariate and its expected value under the model, whereas $\sum_{i=1}^n r_i^2(0)/n$ can be viewed as the average discrepancy without a model. This consideration led O'Quigley and Flandre (1994) to define

$$R^2(\beta) = 1 - \frac{\sum r_i^2(\beta)}{\sum r_i^2(0)} \tag{13.6}$$

This is then a clear analogue to that of R^2 for linear regression. That of itself would not be enough since there may be other possible generalizations. We need study its properties and show that an interpretation for the population equivalent in terms of explained variation holds.

Investigating the impact of censoring

The effect of censoring for large samples on $R^2(\beta)$ was studied by O'Quigley and Flandre (1994) and is so small that it can be ignored in practice, even for rates of censoring between ninety to ninety nine

percent. However, if we are to obtain exact asymptotic results, in which our estimator converges to a quantity unaffected by an independent censoring mechanism, then we need to do a little extra work. This work amounts to weighting 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 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 Kaplan-Meier curve. In practice, ties, if they exist, are split randomly. We then define the quantity $\mathcal{I}(b)$ for $b = 0, \beta$ by

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty \{Z_i(t) - \mathcal{E}_b(Z|t)\}^2 d\hat{F}(t)$$

or, in the more familiar counting process notation by,

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty W(t)\{Z_i(t) - \mathcal{E}_b(Z|t)\}^2 dN_i(t) = \sum_{i=1}^{n} \delta_i W(X_i) r_i^2(b). \tag{13.7}$$

These quantities are, as before, averages of squared residuals, under the null model and under the best fitting model, the only difference being that the average here is weighted with respect to the increments $d\hat{F}(t)$. For large samples we will be able to assert that $\hat{F}(t)$ will be close to $F(t)$ and so our average is taken over time. With this in mind we then appeal to a broadened definition for R^2 in which:

$$R^2(\beta) = 1 - \frac{\sum_{i=1}^n \delta_i W(X_i) r_i^2(\beta)}{\sum_{i=1}^n \delta_i W(X_i) r_i^2(0)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)}. \tag{13.8}$$

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 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.

Population parameter Ω^2

The population parameter $\Omega^2(\beta)$ of $R^2(\hat{\beta})$ was given in O'Quigley & Flandre (1994). $R^2(\hat{\beta})$ can be considered a semi-parametric estimate of $\Omega^2(\beta)$ in as much as it is unaffected by monotonic increasing transformations on time (see Section 3.9). We will see that $\Omega^2(\beta)$ is unaffected by an independent censorship mechanism. If in addition Z is time-invariant, we also see that

$$\Omega^2(\beta) = 1 - \frac{E\{E[Z - E(Z|\mathcal{A}(T))]^2\}}{E\{E[Z - E(Z|\mathcal{B}(T)]^2\}}, \tag{13.9}$$

where $\mathcal{A}(t) = \{t\}$ and $\mathcal{B}(t) = \{u : u \geq t\}$ so that, in view of equation (3.32), $\Omega^2(\beta)$ has the interpretation of the proportion of explained variation. This of itself would not be interesting enough and we also show that this choice of $\mathcal{B}$ is a sensible one. In fact, the results for the above choice, chosen to accommodate sequential conditioning on the risk sets, are very close to those arising under the definition $\mathcal{B}(t) = \mathcal{T}$ (see Table 13.1). Indeed, for practical purposes of interpretability we can take $\Omega^2(\beta)$ to be defined as in the following equation where the approximation symbol is replaced by an equality symbol, i.e.,

$$\Omega^2(\beta) \approx \frac{\text{Var}\{E(Z|T)\}}{\text{Var}(Z)}.$$

O'Quigley and Flandre showed that $\Omega^2(\beta)$ depends only relatively weakly on different covariate distributions, and values of $\Omega^2(\beta)$ give a good reflection of strength of association as measured by β, tending to 1 for high but plausible values of β. The 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 first two stipulations were that perfect agreement or disagreement should reflect itself in a coefficient of absolute

Table 13.1: Ω^2 as a function of β.

covariate*	c	c	d	c	c	c	d
β	0	0.7	0.7	1.4	2.8	4.2	4.2
$\mathcal{B}(t) = \{u : u \geq t\}$	0.0002	0.0990	0.0979	0.2844	0.5887	0.7577	0.8728
$\mathcal{B}(t) = \mathcal{T}$	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 that described below.

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. In the next section we will show that $\Omega^2(\beta) \to 1$ as $|\beta| \to \infty$ and that it increases with the ability to explain survival rankings by the covariates. Before that, we look at a closely related quantity which turns out to be of use.

Alternative measure $R^2_{\mathcal{E}}$

For mostly theoretical purposes we also consider an alternative definition to R^2, in which we use the expected (with respect to the π's) rather than the observed squared residuals. Consider then

$$\begin{aligned}\mathcal{J}(\beta, b) &= \int_0^\infty W(t) \sum_{j=1}^n \pi_j(\beta, t)\{Z_j(t) - \mathcal{E}_b(Z|t)\}^2 d\bar{N}(t) \\ &= \sum_{i=1}^n \delta_i W(X_i) \mathcal{E}_\beta\{r_i^2(b)|X_i\}\end{aligned}$$

and define

$$R^2_{\mathcal{E}}(\beta) = 1 - \frac{\sum_{i=1}^n \delta_i W(X_i)\mathcal{E}_\beta\{r_i^2(\beta)|X_i\}}{\sum_{i=1}^n \delta_i W(X_i)\mathcal{E}_\beta\{r_i^2(0)|X_i\}} = 1 - \frac{\mathcal{J}(\beta,\beta)}{\mathcal{J}(\beta,0)}. \quad (13.10)$$

Our experience indicates that when the proportional hazards model correctly generates the data, $R^2_{\mathcal{E}}$ will be very close in value to R^2. Indeed we will show, under the model, that $|R^2(\hat{\beta}) - R^2_{\mathcal{E}}(\hat{\beta})|$ converges to zero in probability. This coefficient is of interest in its own right although our main purpose here is to use it for developing properties of the next section. It can also be used to construct confidence intervals for the population quantity $\Omega^2(\beta)$, intervals which have, for increasing sample size, exactly the same coverage properties of those for $\hat{\beta}$ itself. Another angle to understand $\mathcal{J}(\beta, b)$ follows from taking the expectation of $\mathcal{I}(b)$ under the model, using the results for counting processes (see for example Fleming and Harrington 1991) we have

$$E\{\mathcal{I}(b)\} = \sum_{i=1}^n \int_0^\infty E\{W(t)[Z_i(t) - \mathcal{E}_b(Z|t)]^2 Y_i(t) \exp[\beta Z_i(t)]\} d\Lambda_0(t),$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. If we replace the unknown Λ_0 by the Nelson-Aalen estimate (Breslow 1972, 1974) and the expectations under the integral by the observed quantities, then we recover $\mathcal{J}(\beta, b)$ as an estimate of $E\{\mathcal{I}(b)\}$. It is also straightforward to verify that $\mathcal{J}(\beta, \beta)$ is the weighted information of the Cox model.

13.5 Finite sample properties of R^2 and $R^2_{\mathcal{E}}$

We have the following immediate lemmas:

Lemma 13.1 *Viewing R^2 as a function of β then: $R^2(0) = 0$ and $R^2(\beta) \leq 1$.*

Lemma 13.2 *$R^2(\beta)$ is invariant under linear transformations of Z and monotonically increasing transformations of T.*

The following lemma is not a precise result, although we have a precise equivalent for large samples. It provides some insight into R^2, viewed as a function of β. It also indicates why, apart from theoretical interest, only $R^2(\hat{\beta})$ need concern us.

Lemma 13.3 *$R^2(\beta)$ as a function of β, reaches its maximum around $\hat{\beta}$.*

Proofs of the above are similar to those given by O'Quigley and Flandre (1994). More details are provided in the chapter on proofs. Note that R^2, unlike $R^2_{\mathcal{E}}$ and Ω^2, cannot be guaranteed to be non-negative. A negative value for R^2 is nonetheless difficult to obtain in practice, corresponding to the unusual case where the best fitting model, in a least squares sense, provides a poorer fit than the null model. $R^2(\hat{\beta})$ will only be slightly negative in such cases if $\hat{\beta}$ is very close to zero.

Lemma 13.4 *An approximate sums of squares decomposition holds for r_i^2 and holds exactly in the following expression:*

$$\mathcal{E}_\beta\{r_i^2(0)|X_i\} = \mathcal{E}_\beta\{r_i^2(\beta)|X_i\} + \{\mathcal{E}_\beta(Z|X_i) - \mathcal{E}_0(Z|X_i)\}^2. \quad (13.1)$$

Both the approximate and the exact sum of squares decomposition, outlined in more detail below, are valuable in underlining the great similarity between proportional hazards models and linear models. Although we do not pursue the idea it would be quite possible to develop for the proportional hazards model a whole theory for testing and fit based on sums of squares and analysis of variance type decompositions. Even F-tests can be constructed, although, at the present time, there appears to be no obvious advantage to any such alternative approach. One consequence of the above breakdown is:

Lemma 13.5 *The coefficient $R^2_{\mathcal{E}}(\beta)$ can be reexpressed as:*

$$R^2_{\mathcal{E}}(\beta) = \frac{\sum_{i=1}^n \delta_i W(X_i)\{\mathcal{E}_\beta(Z|X_i) - \mathcal{E}_0(Z|X_i)\}^2}{\sum_{i=1}^n \delta_i W(X_i)\mathcal{E}_\beta\{r_i^2(0)|X_i\}}.$$

The re-expression of $R^2_{\mathcal{E}}(\beta)$ in the lemma is helpful in obtaining the further lemmas:

Lemma 13.6 *As a function of β, $0 \leq R^2_{\mathcal{E}}(\beta) \leq 1$, and $R^2_{\mathcal{E}}(0) = 0$.*

Whereas $R^2(\beta)$ depends on the observations directly, $R^2_{\mathcal{E}}(\beta)$ is a function of expectations across the observations and although, at least for correctly specified models, there will be close agreement between the $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ (a result made more precise below), the two coefficients behave very differently when viewed as functions of β. In particular, in contrast to Lemma 13.3, we have:

Lemma 13.7 *As $|\beta| \to \infty$ then $R^2_{\mathcal{E}}(\beta) \to 1$.*

We also have:

Lemma 13.8 *$R^2_{\mathcal{E}}(\beta)$ is invariant under linear transformations of Z and monotonically increasing transformations of T.*

The proof of the linearity property follows in the same way as for R^2 (O'Quigley and Flandre 1994), and an outline of the proof of monotonicity is given in the chapter on proofs. The figure helps illustrate the contrasting behaviors of the two coefficients, seen as functions of β. It is clear that $R^2(\beta)$ as a function of β does not increase to 1 as $|\beta| \to \infty$, but rather reaches its maximum near $\hat{\beta}$. The monotonicity property of $R^2_{\mathcal{E}}(\beta)$ also has an interesting connection to the literature on the efficiency of the Cox model, which has also noted that the information $\mathcal{J}(\beta, \beta) \to 0$ as $|\beta| \to \infty$ (Efron 1977, Oakes 1977, Kalbfleisch and Prentice 1980 Section 4.7).

13.6 Large sample properties

The most straightforward approach is to define the population parameter $\Omega^2(\beta)$ as the probability limit of $R^2_{\mathcal{E}}(\beta)$ as $n \to \infty$. We can then investigate separately how meaningful is $\Omega^2(\beta)$, in particular how it can be viewed as an index of explained variation. We then need to

show that $R^2(\hat{\beta})$ converges in probability to $\Omega^2(\beta_0)$ where β_0 is the "true" value under which the data are generated. Let

$$S^{(r)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Y_i(t)e^{\beta Z_i(t)}Z_i(t)^r, \quad s^{(r)}(\beta,t) = ES^{(r)}(\beta,t),$$

for $r = 0, 1, 2, 3, 4$. We assume that the Andersen-Gill conditions hold. First it is straightforward to establish that: $\mathcal{E}_\beta(Z|t) = S^{(1)}(\beta,t)/S^{(0)}(\beta,t)$. Next we have:

Lemma 13.9 *The coefficient $\mathcal{J}(\beta,b)$ can be reexpressed as:*

$$\mathcal{J}(\beta,b) = \int W(t)\left\{\frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - 2\frac{S^{(1)}(\beta,t)S^{(1)}(b,t)}{S^{(0)}(\beta,t)S^{(0)}(b,t)} + \frac{S^{(1)}(b,t)^2}{S^{(0)}(b,t)^2}\right\}d\bar{N}(t).$$

Theorem 13.1 *As $n \to \infty$ $\mathcal{J}(\beta,b)$ converges in probability to $J(\beta,b)$ where*

$$J(\beta,b) = \int w(t)\left\{\frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} - 2\frac{s^{(1)}(\beta,t)s^{(1)}(b,t)}{s^{(0)}(\beta,t)s^{(0)}(b,t)} + \frac{s^{(1)}(b,t)^2}{s^{(0)}(b,t)^2}\right\} s^{(0)}(\beta,t)\lambda_0(t)dt$$

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

The value to which $R^2_{\mathcal{E}}(\beta)$ converges for large samples, i.e.,

$$R^2_{\mathcal{E}}(\beta) \xrightarrow{\mathrm{P}} 1 - \frac{J(\beta,\beta)}{J(\beta,0)}, \tag{13.2}$$

leads to a natural definition for the relevant population parameter via:

Definition 13.2 *Let us take*

$$\Omega^2(\beta) = 1 - \frac{J(\beta,\beta)}{J(\beta,0)}, \tag{13.3}$$

and, from this, we obtain the important convergence in probability result:

Theorem 13.2 $|R^2_{\mathcal{E}}(\beta) - \Omega^2(\beta)| \xrightarrow{P} 0$. *In particular, $\mathcal{J}(\beta,\beta)$ and $\mathcal{J}(\beta,0)$ converge in probability to $J(\beta,\beta)$ and $J(\beta,0)$, respectively.*

Corollary 13.1 $0 \leq \Omega^2(\beta) \leq 1$, $\Omega^2(0) = 0$, *and as* $|\beta| \to \infty$, $\Omega^2(\beta) \to 1$. *Additionally* $\Omega^2(\beta)$ *is invariant under linear transformations of* Z *and monotonically increasing transformations of* T.

We now show that $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ are asymptotically equivalent; therefore $R^2(\hat{\beta})$ is consistent for $\Omega^2(\beta_0)$.

Theorem 13.3 *Under the Andersen-Gill conditions,* $|R^2(\hat{\beta}) - R^2_{\mathcal{E}}(\hat{\beta})| \xrightarrow{P} 0$.

In our own practical experience, when the proportional hazards model holds, there is very close agreement between the coefficients $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ (see the examples below). When discrepancies arise, this is indicative of a failure in model assumptions. We also have that:

Corollary 13.2 $R^2(\hat{\beta})$ *consistently estimates* $\Omega^2(\beta_0)$. *In particular,* $\mathcal{I}(\hat{\beta})$ *and* $\mathcal{I}(0)$ *consistently estimate* $J(\beta_0, \beta_0)$ *and* $J(\beta_0, 0)$, *respectively.*

Theorem 13.4 $R^2(\hat{\beta})$ *and* $R^2_{\mathcal{E}}(\hat{\beta})$ *are asymptotically normal.*

Monotonicity of Ω^2

As strength of association increases so should the measure of correlation or explained variation. We know, from the results of Section 3.9 that Ω^2 is quantifying predictability. We can obtain further insights into this by considering additional properties of Ω^2. For instance we have that increasing strength of association manifests itself via an increasing $|\beta_0|$, once the covariate scale has been fixed. We have

Theorem 13.5 $\Omega^2(\beta_0)$ *as a function of* β_0, *increases with* $|\beta_0|$.

In fact, we will show below that Ω^2 increases with the predictability of survival rankings, which corresponds to Kendall's third stipulation (in the context of the semiparametric Cox regression). Let $Z_j > Z_i$ be the covariates for two subjects in the study, and assume $\beta_0 > 0$ without loss of generality. We can transform all the survival times to exponentially distributed via the transformation $\Lambda_0(\cdot)$, where Λ_0 is the baseline cumulative hazard function. Such a transformation preserves the ranking of the failures so that $\Omega^2(\beta_0)$ is unchanged. Then conditional, on the covariates, a simple calculation shows that

$$\Pr(T_i > T_j) = \frac{\exp(\beta_0 Z_j)}{\exp(\beta_0 Z_i) + \exp(\beta_0 Z_j)},$$

which increases strictly with β_0.

From the above we see that, given the covariates, as the predictability of the survival rankings increases, so does Ω^2. Furthermore, as a result of Theorem 13.5, we can obtain confidence intervals of $\Omega^2(\beta_0)$ from those for β_0, since $\Omega^2(\beta_0)$ is an increasing function of $|\beta_0|$. Only the absolute value conveys information concerning strength of effect and we can then simply invert intervals for β_0, obtained by the usual methods, into intervals for $\Omega^2(\beta_0)$. The coverage properties will then be the same as those already established for the log relative-risk estimate. Since $R^2_{\mathcal{E}}(\beta)$ is consistent for $\Omega^2(\beta)$ for any β then, in practice, we only need to "plug" the two endpoints of the β-confidence interval into $R^2_{\mathcal{E}}$. This gives an approximate confidence interval for $\Omega^2(\beta_0)$. We have not carried out detailed investigation of the coverage properties of such intervals, but in the examples below, we see that such "plug-in" method gives a confidence interval that agrees very well with inference based on bootstrap resampling.

Independent censoring

Here, we assume that C is independent of T and Z. An important property is that the population parameter $\Omega^2(\beta)$ be not affected by the censorship. In order to show this, it helps to recall our earlier discussion on the two roles that time plays in the model. First, $Z(\cdot)$ in general is a stochastic process with respect to time, meaning that $Z(t)$ is a random variable at any fixed t and may have different distributions at different times t. Secondly, the failure time variable T is a non-negative random variable denoting time. While it is immediate to understand the distribution of T given the covariates, we have at any fixed time t two different conditional distributions of $Z(t)$ on T that are of interest to us. One is conditioning on $T \geq t$ under the independent censoring assumption this can be interpreted as given all the subjects that have survived at least until time t and can be estimated by the empirical distribution of $Z(t)$ in the risk set at time t.

Another kind of conditional distribution of interest is that of $Z(t)$ given $T = t$. Under the assumption that T has a continuous distribution we usually observe only one failure at a time and it is difficult to estimate this latter conditional distribution based on a single observation, or a few in the case of ties. We can, however, obtain a consistent estimate by leaning on the model and the main theorem of proportional hazards regression of Section 7.4, one of whose corollaries is: under the model and an independent censorship, the conditional distribution function of $Z(t)$ given $T = t$ is consistently estimated by

$$\hat{F}_t(z|t) \;=\; \hat{P}(Z(t) \leq z|T=t) \;=\; \sum_{\{j:Z_j(t)\leq z\}} \pi_j(\hat{\beta},t).$$

Note that the corollary also applies to multiple dimensional covariates. As a consequence, we also have that:

Corollary 13.3

$$\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)} = E_\beta\{Z(t)|t\}, \quad \frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} = E_\beta\{Z(t)^2|t\}, \quad \frac{s^{(1)}(0,t)}{s^{(0)}(0,t)} = E_0\{Z(t)|t\}.$$

Corollary 13.4 *The cumulative distribution for T can be expressed as*

$$F(t) = \int_0^t w(t)s^{(0)}(\beta,t)\lambda_0(t)dt.$$

Lemma 13.10 *For b in $J(\beta,b)$ taking the values β, 0:*

$$J(\beta,b) = \int E_\beta\{[Z(t) - E_b(Z(t)|t)]^2|t\}dF(t).$$

Corollary 13.5 *We can now rewrite $\Omega^2(\beta)$ as:*

$$\Omega^2(\beta) = 1 - \frac{\int E_\beta\{[Z(t) - E_\beta(Z(t)|t)]^2|t\}dF(t)}{\int E_\beta\{[Z(t) - E_0(Z(t)|t)]^2|t\}dF(t)}. \tag{13.4}$$

We can deduce from the corollary that $\Omega^2(\beta)$ does not involve the censoring distribution. It is therefore unaffected by changes in any independent censoring mechanism, in particular its removal as a mechanism impacting our ability to make observations on T.

13.7 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 section.

A sum of squares decomposition

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

$$\begin{aligned}
&\sum_{i=1}^n \delta_i W(X_i) r_i^2(0) \\
&\quad = \sum_{i=1}^n \delta_i W(X_i) r_i^2(\beta) + \sum_{i=1}^n \delta_i W(X_i) \{\mathcal{E}_\beta(Z|X_i) - \mathcal{E}_0(Z|X_i)\}^2 \\
&\qquad + 2 \sum_{i=1}^n \delta_i W(X_i) \{\mathcal{E}_\beta(Z|X_i) - \mathcal{E}_0(Z|X_i)\} \{Z_i(X_i) - \mathcal{E}_\beta(Z|X_i)\}.
\end{aligned}$$

The last term in the above is a weighted score and therefore converges asymptotically to zero. It is this result which will enable us to break down the total sum of squares into two components: a residual sum of squares and a regression sum of squares. To make this precise we introduce the following definition which, immediately, can be seen to be analogous to those with which we are familiar from ordinary linear regression.

Definition 13.3 *The total, residual, and regression sum of squares are defined by:*

$$SS_{\text{reg}} = \sum_{i=1}^n \delta_i W(X_i) \{\mathcal{E}_{\hat{\beta}}(Z|X_i) - \mathcal{E}_0(Z|X_i)\}^2$$

$$SS_{\text{tot}} = \sum_{i=1}^n \delta_i W(X_i) r_i^2(0), \qquad SS_{\text{res}} = \sum_{i=1}^n \delta_i W(X_i) r_i^2(\hat{\beta}).$$

From this definition we obtain an asymptotic decomposition of the total sum of squares into the residual sum of squares and the regression sum of squares, i.e.

Lemma 13.11 *Asymptotically, the above three quantities are related by:*

$$SS_{\text{tot}} = SS_{\text{res}} + SS_{\text{reg}}. \tag{13.5}$$

We can then conclude that R^2 is asymptotically equivalent to the ratio of the regression sum of squares to the total sum of squares. Notice that for $R^2_{\mathcal{E}}(\beta)$, even with finite samples, we have an exact decomposition of the sum of the squares. Therefore $R^2_{\mathcal{E}}$ can be expressed exactly as the ratio of a 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. In this case, $Z(t) \equiv Z$ and, letting $\mathcal{A}(t) = \{t\}$ and $\mathcal{B}(t) = \{u : u \geq t\}$ then we have that:

$$\begin{aligned} J(\beta, \beta) &= E\{E[Z - E(Z|\mathcal{A}(T))]^2\} \\ J(\beta, 0) &= E\{E[Z - E(Z|\mathcal{B}(T))]^2\} \end{aligned}$$

The first equation is immediate and the second follows since $E_0(Z|t) = E_\beta(Z|T > t)$. We can then claim that Ω^2 is indeed a measure of explained variation, the above expressions fitting in precisely with equation (3.32). It is then clear, and backed up further by the simulations of Table 13.1, that

$$\Omega^2(\beta) \approx 1 - \frac{E\{\mathrm{Var}(Z|T)\}}{\mathrm{Var}(Z)} = \frac{\mathrm{Var}\{E(Z|T)\}}{\mathrm{Var}(Z)}. \tag{13.6}$$

What is more, there is nothing to stop us defining explained variation as in the right-hand side of the equation since the marginal distribution of Z and T can be estimated by the empirical and the Kaplan-Meier estimator, while the conditional distribution of Z given $T = t$ by the $\pi_i(\hat{\beta}, t)$. However, it is not clear that there is any advantage to this and we recommend that all calculations be done via the Schoenfeld residuals, evaluated at $\beta = \hat{\beta}$ and $\beta = 0$.

The agreement shown in the table between the different ways of conditioning is rather remarkable. One almost suspects that there may be an actual equality and that the observed differences are simply due to rounding errors. But we have not been able to show as much. The important thing to conclude is that we have a very clear, and precise, interpretation in terms of explained variation.

Explained variation in T given Z

As just described we can interpret our coefficient as an estimate of the variation in Z explained by T. In the context of proportional hazards regression where inference is not impacted by any arbitrary monotonic increasing transformation on T, then the variances and mean squared errors of Z given T are the correct quantities to use in order to quantify predictive strength. This is not immediately intuitive however and it is frequently argued that what is required is a coefficient built around the

variances and mean squared errors of T given Z. In response to that viewpoint, it could be argued that this amounts to wanting to have your cake and eat it, since by making an appeal to the proportional hazards model we are implying that we wish to suppress or ignore the distributional properties of T given Z and that our model (especially in the light of the main theorem of Section 7.4) only describes the conditional distribution of Z given T.

However, at very little cost and effort, we can, if we wish, base our construction on the same quantities we have worked with so far together with an appeal to Bayes rule. This results in a coefficient with an interpretation as the explained variation in T given Z. Recall that for the case of a bivariate normal distribution the two different ways of defining explained variation result in identical population quantities Ω^2. For other distributions (as is the case here) we nonetheless expect that agreement will be strong. This has been the case in our practical experience. We need two quantities: $\mathrm{Var}(T)$ and $E\,\mathrm{Var}(T|Z)$. The first is readily estimated and often we may wish to estimate it by restricting the time interval to have some upper limit. As for $E\,\mathrm{Var}(T|Z)$, note that:

$$E\{\mathrm{Var}(T|Z)\} = \int_{\mathcal{T}}\int_{\mathcal{Z}} \left\{ t - \int_{\mathcal{T}} t dF(t|z) \right\}^2 dF(t|z) dG(z). \quad (13.7)$$

If there is no censoring then consistent estimates for $\Omega^2_T(Z)$ are found by replacing $F(t)$, $G(z)$ and $F(t|z)$ by the empirical estimates $F_n(t)$, $G_n(z)$ and $F_n(t|z)$ to obtain an estimate, let's call it R^2. By virtue of the Helly-Bray theorem R^2 will provide a consistent estimate of Ω^2. Two major problems arise. The first is that, if the dimension of z is high or even continuous, then the estimates $F_n(t|z)$ may be too unreliable to be of practical use. If we wish to appeal to the proportional hazards model then any estimate of $F(t|z)$ will necessarily involve the unspecified $\lambda_0(t)$. Censoring simply adds to the difficulties. However all of these hurdles are readily overcome by a simple appeal to Bayes rule whereby we can write:

$$E\{\mathrm{Var}(T|Z)\} = \int_{\mathcal{T}}\int_{\mathcal{Z}} \left\{ t - \frac{\int_{\mathcal{T}} u g(z|u) dF(u)}{\int_{\mathcal{T}} g(z|u) dF(u)} \right\}^2 dG(z|t) dF(t). \quad (13.8)$$

Consistent estimates for $\Omega^2_T(Z)$ follow if we can consistently estimate the conditional distribution $G(z|t)$ and the marginal distribution $F(t)$. For the marginal distribution of $F(t)$ we have of course the Kaplan-Meier estimate. This makes an assumption of independence between

the censoring and the failure mechanisms. If we wish to make the weaker assumption of conditional independence then, rather than use the Kaplan-Meier estimate, we appeal to the law of total probability and use a weighted combination of within group Kaplan-Meier estimates. In practice, making this relaxing assumption, has a negligible impact on the estimates of Ω^2. It is simpler then to work with an independent censoring assumption. The main theorem of Section 7.4 enables us to replace $g(z|u)$, at each failure point $u = X_i$, by $\pi_i(\beta, X_i)$ as a result of the expression for $\hat{P}(Z(t) \leq z|T = X_i)$ given by Equation 7.6. All of the calculations involve the very same quantities used to construct the coefficient of explained variation in terms of Z given T. The specificity of the model is made use of via the same appeal to the main theorem of Section 7.4. For estimation purposes, all of the integrals in Equation 13.8 reduce to simple sums beginning with the outer integral which, upon replacing $F(t)$ by the stepwise Kaplan-Meier estimate, means that we sum over the observed failure times. The weights will be the step size of the Kaplan-Meier decrement. The empirical cumulative distribution of the $\pi_i(\beta, X_i)$ is also a step function so that, within the outer sum, we also have an inner sum to approximate the integral. There is quite clearly more work to do in order to obtain the coefficient with a direct interpretation as the explained variation in T given Z and, since the results are anticipated to be very close, it is a matter for the user to decide just how important that precise interpretion is.

13.8 Simulation results

It is helpful to recall some simulations comparing the behavior of R^2 with some of the measures mentioned earlier. We make use of some of the results from Table II of Schemper and Stare (1996). In Table 13.2, 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.

As in Schemper and Stare (1996), there were 100 simulation for each entry of the results. In the table, R^2 is the measure proposed here, ρ^2 is the measure of dependence based on information gain (Xu

Table 13.2: A simulated comparison of different measures ($n = 5000$).

$\exp(\beta)$	% censored	Covariate	R^2	ρ^2	ρ^2_W	$\rho^2_{W,A}$	r^2_{pr}	KS
1	0%	c	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
	90%	c	0.002	0.002	0.000	0.000	0.000	0.000
2	0%	c	0.098	0.102	0.096	0.119	0.092	0.101
	50%	c	0.101	0.108	0.089	0.122	0.093	0.088
	90%	c	0.104	0.105	0.103	0.099	0.074	0.015
	0%	d	0.099	0.102	0.113	0.118	0.096	0.095
	50%	d	0.105	0.110	0.114	0.121	0.096	0.089
	90%	d	0.112	0.106	0.125	0.100	0.076	0.016
4	0%	c	0.281	0.295	0.304	0.338	0.272	0.231
	50%	c	0.303	0.334	0.298	0.344	0.274	0.267
	90%	c	0.325	0.340	0.279	0.342	0.278	0.063
16	0%	c	0.586	0.598	0.623	0.664	0.584	0.354
	50%	c	0.623	0.690	0.622	0.668	0.584	0.564
	90%	c	0.703	0.723	0.605	0.670	0.585	0.188
64	0%	c	0.757	0.758	0.785	0.815	0.754	0.397
	50%	c	0.790	0.848	0.790	0.815	0.730	0.717
	90%	c	0.863	0.876	0.763	0.816	0.694	0.321
	0%	d	0.870	0.681	0.777	0.814	0.707	0.319
	50%	d	0.873	0.860	0.776	0.815	0.718	0.861
	90%	d	0.941	0.756	0.795	0.792	0.701	0.135

and O'Quigley 1999). The last four columns are from Schemper and Stare (1996), where ρ^2_W and $\rho^2_{W,A}$ are from Kent and O'Quigley (1988), the measure r^2_{pr} is from Schemper and Kaider (1997) and 'KS' from Korn and Simon (1990) based on quadratic loss. From Table 13.2 we see that overall there is mostly good agreement among these particular coefficients except for KS.

Unlike all the other measures included, the KS measure does not remain invariant to monotone increasing transformation of time. This measure is most useful when the time variable provides more information than just an ordering. There is noticeably close agreement between ρ^2 and R^2 for the majority of the cases. This may have its root in the fact that both measures are semiparametric and calculated using the conditional probability π's. The numerical results for dichotomous covariates with high hazard ratio 64 reflects the fact that for discrete covariates ρ^2 is bounded away from one as $|\beta|$ increases. However, as

discussed in Xu and O'Quigley (1999) as well as Kent (1983), in practice ρ^2 can usually be interpreted without paying special attention to the discreteness of the distribution.

There are most likely theoretical grounds for anticipating some level of agreement among R^2, ρ^2, ρ^2_W and $\rho^2_{W,A}$. Roughly speaking, R^2 has at it base something like a score statistic while the three versions of ρ^2 a likelihood ratio statistic. Large sample agreement for such statistics has been documented and further exploration may shed light on this. The values of r^2_{pr} tend to be slightly lower than these four coefficients, although the strength of association reflected is similar. The measure r^2_{pr} requires more computation than all the other ones in the table because of the multiple imputation technique employed. Some work has been done (Xu and O'Quigley 1999) on establishing the statistical and interpretative properties of ρ^2. Such work remains to be carried out on the other contenders before they could be proposed for routine implementation.

13.9 Extensions

Multiple coefficients

Assume a multicovariate proportional hazards model with β and $Z(t)$ being $p \times 1$ vectors. Under this model, 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. The value 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. Recall that, in the multivariate setting, the main theorem of Section 7.4 provides us with the estimated joint distribution of the covariate vector Z given time. The

most useful way of summarizing this vector is via the linear combination corresponding to the prognostic index. We then proceed very much as for the univariate setting, making the more general definitions of the coefficients.

Definition 13.4 *For the multivariate case we define* $\mathcal{I}(b)$ *as*

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty W(t)\{\eta_i(t) - \beta' \mathcal{E}_b(Z|t)\}^2 dN_i(t). \tag{13.9}$$

Definition 13.5 *For the multivariate case we define* $\mathcal{J}(\beta, b)$ *as*

$$\mathcal{J}(\beta, b) = \int_0^\infty W(t) \sum_{j=1}^{n} \pi_j(\beta, t)\{\eta_j(t) - \beta' \mathcal{E}_b(Z|t)\}^2 d\bar{N}(t). \tag{13.10}$$

For the univariate case we recover the previous definitions apart from a constant multiple which will cancel. We then have:

Definition 13.6

$$R^2(\beta) = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)}\,; \qquad R^2_{\mathcal{E}}(\beta) = 1 - \frac{\mathcal{J}(\beta, \beta)}{\mathcal{J}(\beta, 0)}. \tag{13.11}$$

Definition 13.7 *In order to describe probability limits we define* $J(\beta, b)$ *to equal*

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

where $a^{\otimes 2} = aa'$ *and* $a \otimes b = ab'$ *for vectors* a *and* b.

The definition leads to:

Lemma 13.12 *Under the Andersen-Gill conditions; letting* $n \to \infty$, *we have*

$$\Omega^2(\beta) = 1 - \frac{J(\beta, \beta)}{J(\beta, 0)}, . \tag{13.12}$$

Notice that although $R^2(\beta)$ and $R^2_{\mathcal{E}}(\beta)$ are not defined for $\beta = 0$, the limits exist and are equal to zero as $\beta \to 0$. So we can define $R^2(0) = R^2_{\mathcal{E}}(0) = \Omega^2(0) = 0$. As in the one-dimensional case, we have the following similar properties:

Theorem 13.6 $|R^2_{\mathcal{E}}(\beta) - \Omega^2(\beta)| \xrightarrow{P} 0$. *In particular,* $\mathcal{J}(\beta, \beta)$ *and* $\mathcal{J}(\beta, 0)$ *converges in probability to* $J(\beta, \beta)$ *and* $J(\beta, 0)$*, respectively.*

Corollary 13.6 $0 \leq \Omega^2(\beta) \leq 1$, $\Omega^2(0) = 0$, *and as* $|\beta| \to \infty$, $\Omega^2(\beta) \to 1$. *Additionally* $\Omega^2(\beta)$ *is invariant under linear transformations of* Z *and monotonically increasing transformations of* T.

We have that $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ are asymptotically equivalent, therefore $R^2(\hat{\beta})$ is consistent for $\Omega^2(\beta_0)$.

Theorem 13.7 *Under the Andersen-Gill conditions,* $|R^2(\hat{\beta}) - R^2_{\mathcal{E}}(\hat{\beta})| \xrightarrow{P} 0$.

In our own practical experience, when the proportional hazards model holds, there is very close agreement between the coefficients $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ (see the examples below). When discrepancies arise, this would seem to be indicative of a failure in model assumptions. We can also see that

Corollary 13.7 $R^2(\hat{\beta})$ *consistently estimates* $\Omega^2(\beta_0)$. *In particular,* $\mathcal{I}(\hat{\beta})$ *and* $\mathcal{I}(0)$ *consistently estimate* $J(\beta_0, \beta_0)$ *and* $J(\beta_0, 0)$*, respectively.*

Theorem 13.8 $R^2(\hat{\beta})$ *and* $R^2_{\mathcal{E}}(\hat{\beta})$ *are asymptotically normal.*

Lemma 13.13 *All three quantities;* $R^2(\beta)$, $R^2_{\mathcal{E}}(\beta)$ *and* $\Omega^2(\beta)$ *are invariant under linear transformations of* Z *and monotonically increasing transformations of* T.

Finally, a sum of squares decomposition can be obtained for both R^2 and $R^2_{\mathcal{E}}$, in the same way as in the one-dimensional case.

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

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

The above coefficient, motivated by an analagous 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 $R^2_{\mathcal{E}}$ and partial Ω^2 can be defined in a similar way.

We can also derive the above definition 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 the above definition.

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 basic definitions 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 ith subject in stratum s. Thus we have

$$r_i(b;s)=Z_{is}(X_{is})-\mathcal{E}_b(Z|X_{is}) \tag{13.13}$$

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

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

From this we can define

$$R^2(\beta) = 1 - \frac{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(\beta, s)}{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(0, s)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)}\,. \qquad (13.14)$$

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 on 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 could be used instead. We would not anticipate this having a great impact on the calculated value of $R^2(\beta)$ but this has yet to be studied.

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 (13.8). 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 and Gustafson 1998). 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 use the sieves method which estimates $\beta(t)$ 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

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

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 $\beta(t)$ has recently been explored in Xu and Adak (2000).

13.10 Theoretical construction for distance measures

The distance measures described in Section 13.3 were defined empirically with no population model in mind. However, it is quite straightforward to set up a theoretical structure enabling ready conclusions concerning large sample behavior (O'Quigley, Flandre and Reiner 1999; Schemper and Henderson 2000). The simulation results of Schemper (1990) are confirmed. Also we will see that the measures can be expected to have an upper bound less than 1 as hinted at by Schemper's empirical investigation and that, for example, in the uncensored case the measures V_1 and V_2 estimate the same population quantity. The theoretical setting makes it clear that, unless further modification is undertaken, the population equivalents of the distance measures are affected by censoring, whether or not independent of the failure mechanism.

Uncensored case

As usual we define the empirical distribution function of survival by $F_n(t)$, the empirical survival distribution conditional on the covariate z by $F_n(t|z)$ and the empirical distribution of the covariate z by $H_n(z)$. Also, we have $S_n(t) = 1 - F_n(t)$ and $S_n(t|z) = 1 - F_n(t|z)$. Finally the individual observations can be re-expressed via the function $Y_t(u)$ where $Y_t(u)$ takes the value 1 when $0 < u < t$, the value 0.5 when $u = t$ and the value 0 otherwise. We keep the definition $Y_t(u) = 0.5$ at the value $u = t$ in order to facilitate comparison with Schemper's work (1990, 1992, 2000). However, as far as large sample theory is concerned,

we could define this to be either zero or one at $u = t$ without impacting population quantities.

Referring to Section 13.3, observe that for an observed survival time t, the function $Y_t(u)$ corresponds to the empirical survival function S_{ij} in which the ith subject fails at time t and the argument u is given values corresponding to the observed failure times.

Note that the inner and outer sums contain n elements where n is the number of independently observed survival times. We have $k_i = n\,,\ \forall i$. Things become more transparent when we multiply the outer sum by n^{-1} in both numerator and denominator. The weak law of large numbers then indicates that these quantities converge in probability to expectations. For the inner sum for example $k_i^{-1} \sum |S_{ij} - \bar{S}_j|^\ell\,,\ \ell = 1, 2$ converges in probability, as $k_i (= n) \to \infty$, to the mean absolute ($\ell = 1$) or quadratic ($\ell = 2$) distance between the marginal survival curve at point u and a randomly chosen subject's empirical curve $Y_t(u)$. This mean is calculated over all possible values of u i.e., with respect to the marginal density of survival. Analogously $k_i^{-1} \sum |S_{ij} - \bar{S}_{ij}|^\ell,\ \ell = 1, 2$ converges in probability to a distance between the conditional survival curves (given the covariate) and $Y_t(u)$, once again over all values of u. The outer sums, multiplied by n^{-1} also converge to expectations. In the uncensored case inner and outer expectations are with respect to the same density, that governing survival. It is then natural to have:

Definition 13.8 *The population quantity θ_ℓ is expressed via the ratio of a denominator D_ℓ and a numerator N_ℓ so that $\theta_\ell = 1 - D_\ell^{-1} N_\ell$ where we write:*

$$N_\ell = \int\int\int |Y_t(u) - S(u|z)|^\ell dF(u) dF(t|z) dH(z), \quad (13.15)$$

$$D_\ell = \int\int |Y_t(u) - S(u)|^\ell dF(u) dF(t). \quad (13.16)$$

The simplest situation in which we can readily see how to obtain a consistent estimate of θ_ℓ arises when z takes a small number of finite values. For each value, we can consider the corresponding empirical quantities: $F_n(t|z)$, $S_n(t|z)$, $F_n(t)$ and $H_n(t)$ and then, in the above equation, we can replace the population quantities; $F(t)$, $F(t|z)$ and $H(z)$ by $F_n(t)$, $F_n(t|z)$ and $H_n(z)$ respectively. We can denote such an estimate by $\hat{\theta}_\ell$ and conclude that it is a consistent estimate for θ_ℓ (O'Quigley, Flandre and Reiner 1999). The consistency follows

from standard results for weak convergence (see Section 3.3) whereby $F_n(t) \to F(t)$, $F_n(t|z) \to F(t|z)$ and $H_n(t) \to H(t)$ at all continuity points t of $F(t)$, $F(t|z)$ and $H(t)$, all arrows indicating convergence in probability.

The above expressions make no appeal to any model and, as such, can be considered to be completely non parametric. Not forgetting that we are still dealing with the uncensored case, we could nonetheless view $S_n(t|z) = 1 - F_n(t|z)$ as stratified estimates under the Cox model, since, the stratified model has no constraints and the corresponding survival estimates reduce to the usual empirical ones. This is artificial but consider the following; the above arguments only require that our estimates be consistent. If the Cox model is correct then the stratified model (essentially no model for discrete z) and the usual model both produce consistent estimates for $F(t|z)$. Thus, if we were to redefine $\hat{\theta}_\ell$ to be as above but with $\tilde{S}(t|z)$, the estimate based on the Cox model (see Chapter 15), in place of $S_n(t|z)$, then the consistency property is unchanged.

Censored case

For the empirical quantities presented by Schemper (1990, 1992) the sums were taken over both the observed censored and failure times. This appears attractive in that as much as the information as possible is being used. However, as shown by O'Quigley, Flandre and Reiner (1999), and in an analogous demonstration using counting process notation (Schemper and Henderson 2000), the property of consistency is lost. To see this we deal separately with the sums of censored observations and those that are uncensored. The quantities denoted k_i still count the number of terms in the respective sums so that we can again make a simple appeal to the weak law of large numbers. The standardized "censored" sum converges to an expectation taken with respect to the density $f_{U|U<t}(u|U < t)$, the conditional density of failure time U given that it is less than t. The standardized "uncensored" sum converges to an expectation taken with respect to $f(u)$. The outer sums concern all observations so that the expectations to which these standardized sums converge is taken with respect to the distribution of the minimum of observed survival and censoring times. The survival distribution for censoring, denoted $G(u)$, though enters explicitly into the calculations. The denominator converges to the sum of two terms: an "uncensored" term and a "censored" term which we can write (O'Quigley, Flandre and Reiner 1999) as:

$$\int\int |Y_t(u) - S(u)|^\ell f(u)G(t)dudF(t)$$
$$+ \int\int_0^t (1 - F(t))^{-1}|Y_t(u) - S(u)|^\ell f(u)dudG(t).$$

The censoring distribution appears explicitly in this expression and any resulting evaluation would be impacted by this distribution. An expression for the numerator can also be worked out (O'Quigley, Flandre and Reiner 1999) and again it involves the unknown censoring mechanism. It would be nice if the censoring distribution were to factor out leading to the property we are aiming for but this is not the case.

Convergence in the censored case

Let us define $\tilde{S}(t)$ to be the usual Kaplan-Meier estimate and $\tilde{S}(t|z)$ to be the proportional hazards estimate of conditional survival, given z. If the model correctly generates the observations, then both $\tilde{S}(t)$ and $\tilde{S}(t|z)$ converge to their population counterparts, $S(t)$ and $S(t|z)$.

Lemma 13.14 *The parameter θ_ℓ is consistently estimated by $\tilde{\theta}_\ell$ where $\tilde{\theta}_\ell = 1 - \tilde{D}_\ell^{-1}\tilde{N}_\ell$ and where $\tilde{N}_\ell$ and $\tilde{D}_\ell$ are defined by:*

$$\tilde{N}_\ell = \int\int\int |Y_t(u) - \tilde{S}(u|z)|^\ell d\tilde{F}(u)d\tilde{F}(t|z)dH_n(z), \tag{13.17}$$

$$\tilde{D}_\ell = \int\int |Y_t(u) - \tilde{S}(u)|^\ell d\tilde{F}(u)d\tilde{F}(t). \tag{13.18}$$

Note that although we have taken $\tilde{F}$ to be the Kaplan-Meier estimator the arguments hold for any other consistent estimator of the true underlying marginal survival curve. Under stronger model assumptions we can work even with a parametric estimator. We might anticipate the Nelson estimator of the survivorship function to produce similar results to those for the Kaplan-Meier estimator. Since $|\hat{\theta}_\ell - \tilde{\theta}_\ell| \to 0$, it follows that $\tilde{\theta}_\ell$ is consistent for θ_ℓ and that, under independent censoring, unlike V_ℓ, it is estimating the same quantity it would have estimated were it possible to remove the censoring. Attempts to extract more information from the censorings, in the absence of further, necessarily strong assumptions, leads to inconsistency if we agree, in this context to take inconsistency to mean that estimators converge to population quantities different to those to which they would have converged were it possible to remove the censoring.

13.11 Isolation method for bias-reduction

In order to motivate this section we first recall the relationship between multiple and partial coefficients which holds in the linear case. When $R^2(Z_{q+1}, \ldots, Z_p|Z_1, \ldots, Z_q)$, is the remaining or partial correlation between the outcome and Z_{q+1} to Z_p after having taken into account the effects of Z_1 to Z_q and $R^2(Z_1, \ldots, Z_p)$ is the multiple correlation with all Z_1 to Z_p in the model, then:

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

This expression holds exactly for the linear model and so, whether we build the multiple correlation by constructing increasingly complex partial coefficients or we define the partial coefficient by increasingly simpler multiple coefficients the final answer is the same. Unfortunately this equation does not hold for other situations which is why there is more than one way to define partial and multiple correlation.

Our suggestion for the multiple coefficient is to reduce it formally to a univariate coefficient via use of the prognostic index. We then defined the partial coefficient via the same expression as the above equation. In any event, whether exact or as an approximation, we can use the equation to make the following simple observation. As we add new variables to the expression for multiple correlation, the value of multiple R^2 will almost certainly increase. Only if the partial correlation for the newly included variable is identically equal to zero will the multiple coefficient stay the same. Sampling error will inevitably lead to squared partial correlations more or less removed from zero and, in turn, for an increasingly biased estimate for the multiple correlation itself. This bias pulls the coefficient in the direction of one and so, in practice, estimated coefficients of explained variation can be quite inflated.

The phenomenon of inflation in the multivariate setting is well known and there are several suggestions for tackling the bias. The most well known remedies are the Akaike Information Criterion, the Bayes Information Criterion, the Schwartz Criterion and Cross-Validation. None of these remedies does very well. For smaller sample sizes they will, typically, over adjust and can even lead to negative squared correlations and, for larger samples, they will mostly not make enough of an adjustment. Apart from Cross-Validation, the scope of these corrections is also very limited and, in the main, is concerned only with

biases due to the dimension of the explanatory variable in relation to the sample size.

In the practical setting of model building the dimension of the covariate vector is only the most immediate and often the least important of several factors which result in inflated estimates. There are indeed many other factors among which: (1) the size of the potential covariate pool from which those used in the model form a subset, (2) the data based transformations on continuous or ordered covariates (3) the stepwise algorithms used to make a selection from the covariate pool, (4) the use of cut-offs to define new derived variables and (5) the inclusion of some relaxation of model assumptions in the light of goodness-of-fit procedures. None of these five factors is usually taken into consideration and yet their impact is far greater than that of the dimension of the final model, in particular when the model has been constructed from a very large data base.

A way which addresses all five factors together with the sixth, the dimension of the covariate vector, is the following. However obtained, a final model is viewed as having two quite distinct underlying construction components. The first of these - the most important in any investigator's eye - is the true strength of effect of the multivariate relationship, however formulated, and which finds its expression in the final model. The second component concerns everything involved in the process which led to writing down that final model. All six of the above factors and any others we may have overlooked are deemed to be a part of this second component.

Let's look a little more closely at this second component. Imagine an investigator who decides to fit a model of dimension five from a data set with one hundred individuals and twenty measured covariates. A second investigator is studying a similar problem on one hundred entirely comparable individuals but this time, instead of twenty covariates to choose from, he has two hundred. A third investigator finds him or herself in a situation comparable to that of the second investigator but has results from two separate data sets. It is clear that the bias here is increasing. It is also difficult to have any idea as to what the size of this bias might be. None of the usual techniques address this form of bias. Next, suppose an investigator decides that all the skew distributions should be subject to log-transformations and, if such a transformation leads to a more significant result then the transformation is maintained, otherwise we leave the scaling as it was. He or she then decides that, for the purposes of interpretation, some continuous

variables will be broken down into categorical variables. If the effect across the ordered categories is comparable, as judged by the regression coefficients, then the $p-1$ binary variables describing p groups are replaced by a single ordinal variable. Note also that, if the spacing of the effect is not the same, it can be made so by rescaling.

A model for true and overfitted effects

There are almost endless ways of fine tuning any model and, in the process, as many ways of inflating our idea as to how predictive the model really is. The kind of transformations just suggested will typically indicate a more predictive model than is really warranted. They are also used very frequently by investigators.

We suppose the following model for these two components: the first being the true strength of effect and the second, everything else involved in the construction of the model. The covariate vector of interest is Z and, as usual, we would calculate $R^2(Z)$, a quantity which we observe. However, we would really like to calculate the multiple correlation given the fitting. We write this as $R^2(Z|F)$ where F is not something we measure, or observe, but is a conceptual quantity indicating the sum total of all the actions taken during the fitting process. We might consider these actions taken on their own in which case we would have $R^2(F)$. The observed multiple $R^2(Z)$ simultaneously involves, as well as the real effects, the fitting process, an important fact made explicit by writing, $R^2(Z) = R^2(Z, F)$. Note that:

$$1 - R^2(Z, F) = [1 - R^2(F)][1 - R^2(Z|F)] \tag{13.19}$$

There are three quantities in the above expression and only one of them, $R^2(Z, F)$, can be observed. If we were able to obtain $R^2(F)$ then the quantity we are really interested in, $R^2(Z|F)$, the true impact of the covariates having removed those effects due to the fitting, becomes immediately available from the above equation.

Estimating the overfitted effects

As a first approximation we can suppose that the fitting effects themselves are orthogonal to the true effects. By this we mean that the amount of inflation, as measured by $R^2(F)$, only depends on the fitting procedures and extraneous factors such as sample size etc. As true effect increases, the population equivalent of $R^2(Z, F)$ will, of course,

also increase but it would not be unreasonable to suppose that the pure inflation factor alone, as measured by $R^2(F)$, depends only weakly, if at all, on any true effect. In particular we will calculate $R^2(F)$ in the absence of any effect and use this value when there are non-zero effects. Recalling the main theorem of Section 7.4 we have that, at each ranked failure time t, the probabilities of choosing individual with covariate vector $Z_i(t)$ obtains from:

$$\pi_i(\beta(t), t) = \frac{Y_i(t)\exp\{\beta(t)Z_i(t)\}}{\sum_{j=1}^{n} Y_j(t)\exp\{\beta(t)Z_j(t)\}}. \tag{13.20}$$

This mechanism is assumed to generate the observations. Suppose that there is no effect. Then the coefficient vector, $\beta(t)$ is identically equal to zero. At each failure time t, letting $n(t) = \sum Y_i(t)$ be the number of subjects in the risk set then, from Equation 13.20, the probability that any individual is chosen is simply $1/n(t)$.

We keep the risk sets fixed, i.e., we condition on the observed risk sets and, from these we sample individuals, each with probability $1/n(t)$ at time point t, thereby establishing a simulated data set in which the true effect is zero. On the basis of these data, the investigator can proceed to use the same fitting procedures, and strategies, that he or she has used on the unmodified data. Stepwise searches, transformations, maximizations, eliminations following goodness-of-fit, categorizations and any other used modeling strategy is replicated on this same data set. For the resulting multivariate model, corresponds an R^2 coefficient. We write this as $R^2(F)$. The more involved, elaborate and exhaustive the fitting technique the higher, on average, we anticipate $R^2(F)$ to be. Overfitting the data manifests itself directly in the coefficient $R^2(F)$.

Some further observations on this whole process are worth making. Firstly, we do not have just a single value of $R^2(F)$. Under a further replication we would, typically, obtain a different value of $R^2(F)$. Under a large number of replications we would obtain a whole, simulated, distribution of values of $R^2(F)$. If we denote by u any one of these replicated values and by $H(u)$ the empirical distribution function of the replications, then we can take $R^2(Z|F)$ by using Equation 13.19 to be:

$$R^2(Z|F) = \int \left(\frac{R^2(Z,F) - u}{1-u}\right) dH(u). \tag{13.21}$$

It can also be interesting to consider the whole distribution of $R^2(Z|F)$, as induced from $H(u)$, rather than just the mean. Another point to note is that, by conditioning on the risk sets, we allow the possibility that in the replications, the same subject could be selected more than once. This may seem odd if we are interpreting being selected as a failure (which indeed it is) but this is only a formal procedure, respecting the probability model which we assume to be generating the observations. That the same subject could not in practice fail more than once is something which does not impact our construction and is in fact required if we do not wish to include complex calculations involving the censoring mechanism. This is not unlike the bootstrap which can also involve repetitions which the design itself could not have produced. Finally, in Equation 13.21 a good approximation would arise from taking the mean $\bar{u}$ across the replications of $R^2(F)$ and writing $R^2(Z|F) \approx [R^2(Z,F) - \bar{u}]/(1-\bar{u})$.

Bias reduction

We call the above the isolation method by which the effects of interest are isolated from those which are artificially generated through the process of model construction. The basic idea is derived from the chaotization principle developed by Kipnis (1977). Kipnis studied the tails of the distribution of an R^2 type measure and how changes in this distribution, occurring by the inclusion of additional factors, could be anticipated by the fitting process alone. His focus was on the significance level of the multivariate coefficient rather than bias reduction itself but the central idea is the same. It requires replication under a model of no association. Kipnis's idea was to use permutation distributions which could be generated under an assumption of no effect whereby all permutations would be considered equally likely. For our particular case, we do not need to carry out any permutation. It is enough to sample based on the probabilities given by Equation 13.20 in which we fix $\beta(t)$ at the value zero.

The approach can lead to significant reduction in bias caused by overfitting, especially when dealing with a large number of covariates. The method is easy to implement and can be adapted readily to deal with more complex situations. For example, we may wish to focus on some factor after having taken account of several factors already included in the model. Here, in order to generate the relevant distribution for $R^2(F)$, and referring to the multivariate version of Equation

13.20, we would fix at zero the coefficient corresponding to the factor of interest and allow those factors for which we are adjusting to be replaced by estimates. These would be constrained estimates in that the coefficient corresponding to the additional factor of interest is always fixed at zero. We then use Equation 13.20 with these values in order to generate the distribution $H(u)$.

13.12 Illustrations from studies in cancer

Study in leukemia

The first example concerns the Freireich (1963) data, which records the remission times of 42 patients with acute leukemia treated by 6-mercaptopurine (6-MP) or placebo. The estimate of the regression coefficient is $\hat{\beta} = 1.53$, and $R^2(\hat{\beta}) = 0.386$ and $R^2_{\mathcal{E}}(\hat{\beta}) = 0.371$. The 95% confidence interval for $\Omega^2(\beta)$, obtained using the monotonicity of $R^2_{\mathcal{E}}(\beta)$, i.e., inverting the interval for β, is (0.106, 0.628). On the basis of 1000 bootstrap samples we find a simple percentile interval as (0.154, 0.714) using R^2, and (0.154, 0.715) using $R^2_{\mathcal{E}}$. The bootstrap mean is 0.413 for R^2 and 0.405 for $R^2_{\mathcal{E}}$, which gives estimated bias of 0.028 and 0.034, respectively. This suggests that bias correction may be necessary, and employing Efron's bias-corrected accelerated bootstrap (BCa) method we have confidence interval (0.111, 0.631) using R^2, and (0.103, 0.614) using $R^2_{\mathcal{E}}$. We see that these have very good agreement with one another (suggesting that the proportional hazards assumption is a reasonable one) as well as the interval obtained through monotonicity.

In Figure 13.1 we plot the values of $R^2(\beta)$ (dots) and $R^2_{\mathcal{E}}(\beta)$ (circles) for the Freireich data versus different values of β. The figure illustrates well the facts that $R^2(\beta)$ reaches a maximum at around $\beta = 1.5$, which is the value of our estimate $\hat{\beta}$ and that $R^2_{\mathcal{E}}(\beta)$ increases with β, approaching 1 as $\beta \to \infty$. Notice that $R^2(\beta) = R^2_{\mathcal{E}}(\beta)$ occurs somewhere between $\beta = 1.5$ and 1.6, again around our estimate $\hat{\beta}$. This is to be anticipated in view of Theorem 13.7.

The above $R^2(\hat{\beta})$ can be compared with some of the other suggestions mentioned in the introduction. For the same data the measure proposed by Kent and O'Quigley (1986) resulted in the value 0.37, and the measure of explained randomness (Xu and O'Quigley 1999), described in the following chapter, obtains the value of 0.40. The explained variation proposals of Schemper (1990), based on empirical

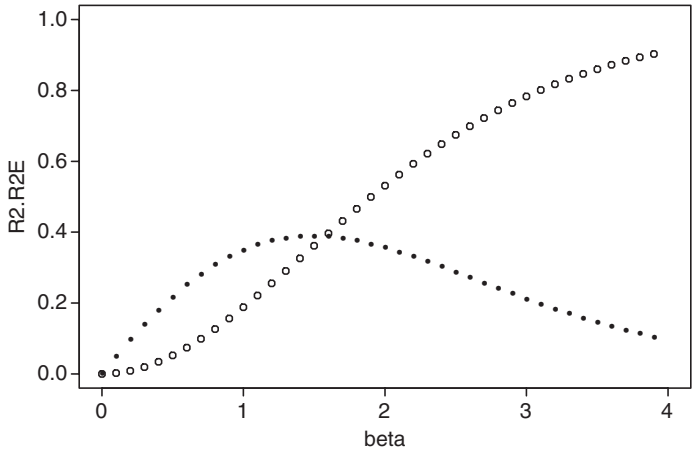

Figure 13.1: A plot of R^2 and $R^2_{\mathcal{E}}$ as functions of β given the observations.

survival functions per subject, resulted in (his notation) $V_1 = 0.20$ and $V_2 = 0.19$ and Schemper's later correction (1994) resulted in $V_2 = 0.29$, although all three of the Schemper measures depend heavily on the censoring (O'Quigley et al. 1999, Schemper and Henderson 2001). The measure of Schemper and Kaider (1997) resulted in $r^2_{pr} = 0.34$. The measure of Korn and Simon (1990), based on quadratic loss, gave the value 0.32. This measure does not remain invariant to monotone increasing transformation of time. For these data the value 0.32 drops to 0.29 if the failure times are replaced by the square roots of the times.

Study in breast cancer

The available data consist of 1504 patients with complete covariate information, among whom there were 357 recorded deaths. The 5 and 10 year survival rates were 0.83 and 0.70, respectively. Of the five covariates, age has a range of 23-55 years, with a median of 45 years. About 6%, 20%, 28%, 27% and 19% of the patients had histology grade 0, 1, 2, 3, and 4, respectively. About 45%, 24%, 23%, 5%, and 2% of the patients had stage 1, 2, 3, 4, and 5 disease, respectively. Out of the 1504 patients, 1075 (71%) had positive progesterone receptor status. The maximum tumor size was 170mm, with a median of 30 mm.

In univariate analysis under the proportional hazard model, all variables are highly significant (Table 13.3). We also calculated the

Table 13.3: R^2 analysis of breast cancer data. Upper part of the table shows results for univariate analyzes. Lower part shows the nested multivariate coefficients.

Single covariate	$\hat{\beta}$	p-value	R^2
Age	-0.24	<0.01	0.005
Histology	0.37	<0.01	0.12
Stage	0.53	<0.01	0.20
Receptor	-0.73	<0.01	0.07
Size	0.02	<0.01	0.18

Covariates in multivariate model	R^2	partial R^2
Age	0.01	
Age and histology	0.12	0.12
Age, histology, and stage	0.26	0.16
Age, histology, stage, and receptor	0.33	0.09
Age, histology, stage, receptor, and size	0.33	0.01

univariate R^2's, and we see that the predictive powers are 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 non-proportional regression effect. This might explain the observed discrepancy between R^2 and $R^2_{\mathcal{E}}$. So we fit a simple two-stage model with the regression effect dropping to zero after a certain change point. When the change point is chosen at 24 months, R^2 from the fitted model turns out to be 0.238, and $R^2_{\mathcal{E}}$ 0.332. 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 the lower part of Table 13.3, we calculated the multiple R^2 for a set of nested models. It also contains the values of the partial R^2 when each additional covariate is added to the existing model. The partial coefficient for tumor size having accounted for the other four variables is only 0.006, suggesting that the extra amount of variation

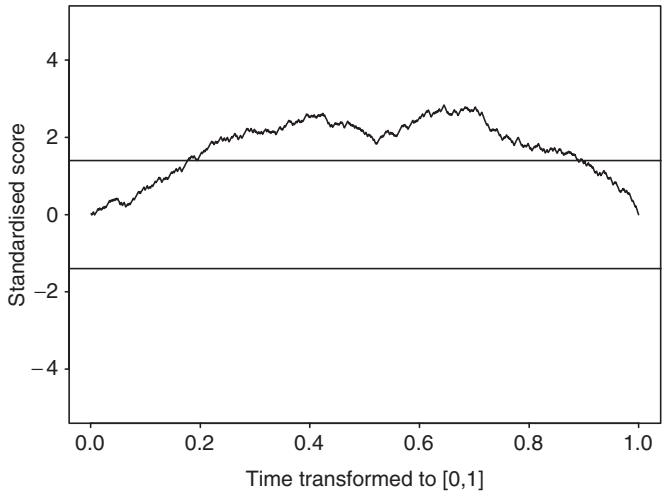

Figure 13.2: *Goodness-of-fit plot for the covariate grade.*

in survival explained by the patient's tumor size is small. Some covariates in this dataset are known to have non-proportional regression effect. Figure 13.2 shows a goodness-of-fit process (see Chapter 8) for the proportional hazards assumption. Recall that the maximum absolute value exceeding the boundary of 1.36 corresponds to the 0.05 significance level of an underlying Brownian bridge when the model is correctly specified. The variation in survival times of these breast cancer patients are mainly explained by three of the five covariates: histology grade, stage, and progesterone receptor status. Xu and Adak (2001) examined closely the time-varying regression effects of these three variables using a tree-based approach. The tree method gives piecewise constant estimated log relative risks. After obtaining a set of nested trees, as one of the methods for selecting a final tree, they used the coefficient R^2_{ex} defined above to arrive at a final tree with two cutpoints at 27 and 46 months. The estimated piecewise-constant regression effects of the three covariates are reproduced in Table 13.4.

While the R^2 from a proportional hazards model with these three covariates is 0.32, the R^2 from the above fitted three piece $\beta(t)$ model is 0.51. For the latter R^2 the calculation used $\hat{\beta}_1(t) \equiv 0$ and $\hat{\beta}_2(t) = \hat{\beta}(t)$. as given in Table 13.4. The improvement in explained variation here reflects also an improvement in fit, underlining the relationship between predictability and goodness-of-fit.

Table 13.4: $\hat{\beta}(t)$ (standard error) from Xu and Adak (2001).

Variable	$t \in [0, 27]$	$t \in [28, 46]$	$t \in [47, 165]$
Histology	0.653 (0.125)	0.362(0.094)	0.201 (0.069)
Stage	0.607 (0.094)	0.349(0.089)	0.365 (0.071)
Receptor	-0.803 (0.225)	-0.708(0.201)	-0.293 (0.164)

Study in gastric cancer

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 (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 logarithmic 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 13.5 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 can then be 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.

Study in multiple myeloma

A further example is motivated by the increasing number of studies carried out in cancer research to correlate the outcome with multi-dimensional molecular and genetic markers. 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; in particular, the Durie-Salmon stage has the smalles R^2 of 0.004. When all 13

Table 13.5: Univariate analysis of gastric cancer data (Rashid et al. 1982).

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

covariates are included in a multivariate Cox model, only six of them remain significant (p-value < 0.08), with the multivariate $R^2 = 0.202$. In particular, the traditional staging system is no longer significantly predictive of survival given the laboratory measurements. Leaving out 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 remaining five that has the highest partial R^2, and so on. The lower part of Table 13.6 gives the nested models and the corresponding R^2's. The data come from a clinical trial (EST 9486) of multiple myeloma conducted by the Eastern Coorperative Oncology Group (Oken et al. 1999). The trial enrolled 653 patients to three randomized arms; VBMCP alone, VBMCP with added HiCy and rIFNα2. No significant survival difference were found across the three arms. The trial collected laboratory measurements on patients' myeloma cells, including measurements from blood or serum: albumin (1 if $\geq$ 3g/l, 0 otherwise), β_2 microglobin 1 if $\geq$ 2.7mg/dl, 0 otherwise), creatinine (1 if $\geq$ 2mg/dl, 0 otherwise), cytoplasmic-immunoglobin heavy chain IgA and IgG (1 if present, 0 absent), kappa light chain (1 if present, 0 absent), percent plasma cells (1 if $\geq$ 0.3%, and hemoglobin (1 if $\geq$ 10g/dl, 0 otherwise); characteristics of circulating myeloma cells: plasma cell labelling index (a measure of plasma cell proliferation, 1 if $\geq$ 1, 0 otherwise), IL-6 receptor status (1 if $\geq$ 270ng/ml, 0 otherwise), and level of C-reative protein (1 if $\geq$ 2ng/ml, 0 otherwise).

All of the above variables, which were originally continuous, were dichotomized using previously published threshold values. 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-

Table 13.6: R^2 analysis of the Myeloma data. Upper part of the table shows results for univariate analyzes. Lower part shows the nested multivariate coefficients.

Single 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

covariates in multivariate model	R^2
PLC	0.08
PLC, creatinine	0.11
PLC, creatinine, plasma	0.13
PLC, creatinine, plasma, a13q	0.16
PLC, creatinine, plasma, a13q, β_2 mcrglb	0.17
PLC, creatinine, plasma, a13q, β_2 mcrglb, IL-6	0.18
All 13 variables	0.20

situ hybridization (FISH) in the laboratory of R. Fonseca at the Mayo Clinic; the corresponding variable a13q was coded 1 if present, 0 absent. We also include the traditional Durie-Salmon stage (1 if I or II, 0 if III) which was routinely used to predict prognosis in multiple myeloma before the availability of assays to measure genetic and other molecular abnormalities of the myeloma cells.

Univariate Cox regression analysis indicates that all of the above 13 covariates are more or less associated with patients' survival times and most of them are highly significant. Table 13.6 shows the estimated regression effects and the standard errors, together with the univariate R^2 coefficients. Effects are not very strong. Even when all 13 variables are included in the analysis we still have an estimated eighty percent of

the variance remaining unexplained. And the true figure would most certainly be higher since no accommodation has been made for the fitting biases.

The same data set was also analyzed by Huang and Harrington (2002), who proposed a penalized partial likelihood approach to the handling of high-dimensional covariates in the proportional hazards regression. The authors pointed out that because there were 270 deaths among the 295 subjects in this data set, the standard partial likelihood estimate in a Cox model with all 13 covariates should be reasonably stable. Even so, in their Table 3 one sees obvious reduction in both the magnitude of the regression effects and the standard errors of the penalized partial likelihood estimate, as compared to the unpenalized estimate. The penalty parameter in their procedure was chosen to minimize a bootstrap estimated mean squared prediction error of the prognostic index. Although the R^2 measure has been so far defined in terms of the usual estimates, it would be straightforward to extend the definition to the penalized partial likelihood estimate. In this case, it turns out that $R^2(\hat{\beta}_\lambda) = 0.198$, almost the same as the $R^2 = 0.202$ with the standard partial likelihood estimate.

Value of R^2 in applied studies

In two of the above examples effects were quite strong and in the other two, although clearly present, effects were relatively weak. This was picked up by the R^2 coefficients and the partial coefficients in particular enable us to decide how practically useful to any prognostic assessment is the inclusion of additional information. Although we have pointed out that R^2 is concerned with prediction and not fit (as often thought) the issues of fit and prediction are not orthogonal to one another. They impact one another in important but different ways. Improving a poor fit will very likely lead to increases in predictive ability as reflected in R^2. A perfect fit (in the sense that the observations are exactly generated by the supposed model) can correspond to an R^2 taking any value between zero and one. A very high value of R^2 can also correspond to a very poor fit. All of that said, in the endeavor to improve our predictive capability, we need consider, alongside one another, both fit and R^2. The fit can be improved by a relaxation of model assumptions, such as the use of a stratified model, or by the introduction of time dependent effects such as the use of changepoint models. Either way it can be worth looking at the plot indicating the

quality of the fit and making sure that we are broadly satisfied with this before presenting our summary R^2 indices of prediction.

13.13 Exercises and class projects

1. A number of suggested coefficients of explained variation, adapted from linear regression, depend on the censoring even when independent of the failure mechanism. Why is this a handicap? Mostly the dependence is such that the higher the censoring the closer to zero is the adapted coefficient. It might be argued that, as the censoring increases, our ability to predict declines and, in consequence so ought a suitable coefficient. Comment on this reasoning.

2. Suppose you are the statistician analyzing the gastric cancer data. The investigating clinician, who has some rudimentary knowledge of statistics, wishes to understand just what you mean by saying that *the value of R^2 for a model with ACT and log(CEA) is 0.37 and this increases to 0.54 when stage is also included. On the other hand the corresponding partial R^2 is equal to 0.27.* How do you answer this question.

3. Using the δ-method and the expression, $1 - R^2(\beta) = \sum r_i^2(\beta) / \sum r_i^2(0)$, derive an approximate confidence interval for $R^2(\hat{\beta})$. For the Freireich data, compare this interval with that obtained in the example on the basis of bootstrap sampling. Comment.

4. Repeat the exercise of the previous question, only applying this time the delta method to $\log[R^2(\hat{\beta})/\{1 - R^2(\hat{\beta})\}]$. What advantages, if any, are there to working with this transformation rather than working with $R^2(\hat{\beta})$ directly?

5. In the broadened definition of $R^2(\beta)$ we have

$$\mathcal{I}(b) = \sum_{i=1}^{n} \int_0^\infty \{Z_i(t) - \mathcal{E}_b(Z|t)\}^2 d\hat{F}(t),$$

where $\hat{F}(t)$ is the Kaplan-Meier estimator. Suppose that $F(t : \theta)$ is a parametric model of the marginal survival curve where θ is a parameter, possibly vector-valued. Investigate a definition of $\mathcal{I}(b)$ in which, instead of $\hat{F}(t)$, we work with $F(t : \hat{\theta})$. What might be the advantages

and drawbacks of such an approach? From the results we already have is it possible to deduce properties such as consistency? If so, under what conditions?

6. For the standard linear model, the coefficient $R(\beta)$, viewed as a function of β is maximized when $\beta = \hat{\beta}$ and where $\hat{\beta}$ is the usual least squares estimate. Thus, for the linear case, a consistent estimator of β obtains by maximizing $R^2(\beta)$. Is this true for the $R^2(\beta)$ defined here for the proportional hazards model? Investigate $\sup_\beta R^2(\beta)$, given data, as an estimate for β. How does it compare with more commonly used estimates?

7. For the normal linear model the transformation $Z = \tanh^{-1}(R)$ has two advantages: the first is that $E(Z)$ provides a very close approximation to $\tanh^{-1}(\Omega)$, the second is that $\mathrm{Var}\,(Z)$ does not depend upon Z, and therefore upon Ω, to a high level of approximation. Furthermore $\mathrm{Var}\,(Z) \approx 1/(n-3)$ where n is the sample size. Discuss this transformation in the context of the $R^2(\beta)$ presented in this chapter. Investigate this more deeply using simulated data.

8. In the previous question, on the basis of simulations, we anticipate that $\mathrm{Var}\,(Z) \approx 1/(n-3)$ where $Z(\beta) = \tanh^{-1}\{\sqrt{R^2(\beta)}\}$. We might conjecture, in the presence of independent censoring, and where k represents the total number of failures, that $\mathrm{Var}\,(Z) \approx 1/(k-3)$. Use simulated data to investigate this assertion. Present an informal argument as to why such a result might hold.

9. We know that, as $|\beta| \to \infty$ then $R^2_{\mathcal{E}}(\beta) \to 1$ and that, as $|\beta| \to \infty$ then $R^2(\beta) \to 0$. This might at first glance appear puzzling. Explain just what is taking place. Explain also why, if the model is correct, we anticipate that $R^2(\hat{\beta})$ and $R^2_{\mathcal{E}}(\hat{\beta})$ will closely agree.

10. Consider an arbitrary proportional hazards model, with unknown cumulative hazard rate $\Lambda_0(t)$, and known regression coefficient vector β_0. For two randomly chosen individuals, the first with covariate vector given by Z_j, the second with covariate vector given by Z_i, show that the probability that the second individual outlasts (survives longer) than the first is given by

$$\Pr(T_i > T_j) = \frac{\exp(\beta_0 Z_j)}{\exp(\beta_0 Z_i) + \exp(\beta_0 Z_j)}.$$

Note that this expression does not involve $\Lambda_0(t)$. How does this result help in the interpretation of R^2.

11. Refer back to Section 3.9 in order to construct a coefficient of partially explained variation from first principles. For a given data set compare this coefficient with that suggested in this chapter based on use of the multivariate coefficient defined in terms of the prognostic index. Show how we could derive an alternative definition of multivariate explained variation based on lower order partially explained variation. Comment on the advantages and disadvantages of this.

12. Suppose for some given data we are considering using an additive risk model or a multiplicative risk model, both of which employ only constant regression coefficients. Consider how we might use an R^2 measure to discriminate between these two models.

13. Using a large data set with a large number of potential risk factors, construct, on the basis of $R^2(Z)$ for some vector Z, as predictive a model as possible, noting down every step made in the construction of the model. Now, on the basis of the isolation method, calculate $R^2(Z|F)$. Compare the sizes of $R^2(Z)$ and $R^2(Z|F)$ and comment on your findings.

14. Generate or use a data set in which only very few observations are censored and in which the covariate is continuously measured. Carry out the usual analysis and calculate R^2. Next, throw away the small percentage of censored observations, replace survival time by $\log T$ and calculate the usual squared product moment correlation coefficient between the covariate and $\log T$. Compare and discuss the results.