The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2006)
6, Number 1, pp. 83–96

Explained variation for survival models

Patrick Royston
Cancer Group
MRC Clinical Trials Unit
222 Euston Rd.
London NW1 2DA, UK
patrick.royston@ctu.mrc.ac.uk

Abstract. This article introduces a new measure of explained variation for use with censored survival data. It is a modified version of a measure previously described by John O'Quigley and colleagues, itself a modification of Nagelkerke's earlier proposal for a general index of determination. I describe Stata programs `str2ph`, which implements the new measure, and `str2d`, which implements a measure proposed in 2004 by Royston and Sauerbrei. I provide examples with real data.

Keywords: st0098, censored survival data, regression models, index of determination, explained variation, explained randomness, information gain

1 Introduction

Data analysts are familiar with the concept of explained variation or index of determination for the linear regression model $y \sim N\left(\mathbf{x}\boldsymbol{\beta}, \sigma^2\right)$ with covariate vector $\mathbf{x}$ and parameter vector $\boldsymbol{\beta}$. The explained variation statistic R^2 may be written less familiarly than usual as

$$R^2 = \frac{\mathrm{var}\left(\mathbf{x}\boldsymbol{\beta}\right)}{\sigma^2 + \mathrm{var}\left(\mathbf{x}\boldsymbol{\beta}\right)} \tag{1}$$

where the variance is taken over the distribution of $\mathbf{x}\boldsymbol{\beta}$ between individuals. Several authors over the last 15 years have proposed versions of explained variation statistics for use with (possibly censored) survival data. These measures have not necessarily been extensions of the linear regression case. For example, Graf et al. (1999) based their proposed measure on the Brier score for survival data, whereas others (e.g., Schemper 1990) have worked with the survival curves fitted from a model.

Survival analysis methodology, particularly the Cox model, is often used to develop so-called prognostic models in diseases such as cancer, where the outcome is time to death or disease recurrence. It is medically relevant to ask how much of the variation in the outcome (survival time) is accounted for through the prognostic index, $\mathbf{x}\boldsymbol{\beta}$. Royston and Sauerbrei (2004) listed desirable characteristics of a measure of discrimination of a survival model, which is closely related to explained variation. Some properties of a good measure include (a) approximate independence of the amount of censoring; (b) reduction to (or a close relationship with) the usual R^2 that would be obtained by an "equivalent" linear regression analysis of the same dataset, if that were possible; (c) the nesting property, that is, for two models $M_1 \subset M_2$ ('$\subset$' denoting nesting) then $R^2(M_1) < R^2(M_2)$; (d) R^2 increasing with the strength of association; and (e) availability of confidence intervals.

Here I will draw together several related lines of previous research. Nagelkerke (1991) proposed a general index of determination that depends on the likelihood-ratio statistic or model χ^2 (minus twice the difference in log likelihoods between the model of interest and the null model) and the number of observations. This index has the major advantages of availability for many model types—including, if partial likelihood is used, the Cox proportional hazards model and its relatives—and of possessing the nesting property, as just defined. Recently, O'Quigley, Xu, and Stare (2005) proposed for use with the Cox model a modified version of the Nagelkerke R^2 that they called ρ_k^2, in which the number of observations is replaced by the number of uncensored observations (events). I shall further modify ρ_k^2, as explained in the next section.

I shall describe a Stata ado-file, `str2ph`, that implements the modified version of ρ_k^2 for survival data. This command is intended for use with proportional hazards (PH) models, that is, with `stcox`, `streg` (PH models—distributions `exponential`, `gompertz`, `weibull`) and Royston's `stpm` routine for flexible parametric survival modeling (Royston 2001, Royston and Parmar 2002) with `scale(hazard)`. The statistic is provided also for other survival models (`streg` with distributions `lnormal`, `llogistic`, `gamma`, and `stpm` with `scale(normal)` and `scale(odds)`), but with a minor caveat that the interpretation of ρ_k^2 is less clear for these distributions. Confidence intervals are provided by using a built-in bootstrap procedure. Also included here, and implemented in the ado-file `str2d`, is the R_D^2 measure of explained variation for survival models described by Royston and Sauerbrei (2004). R_D^2 is a transformation of Royston and Sauerbrei (2004)'s D measure of discrimination of a survival model. I give examples with a real dataset.

2 Measures of explained randomness and explained variation

Nagelkerke (1991)'s general measure of the strength of dependence of the outcome on the predictors in a regression model is defined as

$$\rho_n^2 = 1 - \exp\left\{-\frac{2}{n}\left(l_{\widehat{\boldsymbol{\beta}}} - l_0\right)\right\} = 1 - \exp\left(-\frac{X^2}{n}\right) \qquad (2)$$

where n is the sample size, $l_{\widehat{\boldsymbol{\beta}}}$ denotes the maximized log likelihood of the model, and l_0 is the log likelihood of the comparator (null) model. Suppose that the model has covariate vector $\mathbf{x}$ and linear predictor (index) $\mathbf{x}\boldsymbol{\beta}$. Then $X^2 = 2\left(l_{\widehat{\boldsymbol{\beta}}} - l_0\right)$ is the likelihood ratio statistic for comparing the model with index $\mathbf{x}\boldsymbol{\beta}$ with the null model $\mathbf{x}\boldsymbol{\beta} = 0$. X^2 is distributed as χ^2 on $\dim(\boldsymbol{\beta})$ degrees of freedom under the null hypothesis that $\boldsymbol{\beta} = 0$. It is assumed, where relevant, that $\boldsymbol{\beta}$ excludes the constant, β_0.

In the context of censored survival data, O'Quigley, Xu, and Stare (2005) noted that for a given model and dataset, ρ_n^2 is negatively correlated with the proportion of censored observations and in fact tends to 0 as that proportion tends to 1. For PH models, they proposed replacing the denominator n in (2) by the number of events (uncensored observations), e, to give a new statistic

$$\rho_k^2 = 1 - \exp\left(-\frac{X^2}{e}\right) \tag{3}$$

O'Quigley, Xu, and Stare (2005)'s motivation for (3) was its simplicity together with their demonstration, under certain conditions, of approximate equivalence between ρ_k^2 and several other explained variation–like measures for the PH model, as briefly outlined below. These measures are ρ_W^2 (Kent and O'Quigley 1988), $\rho_{W,A}^2$ (Kent and O'Quigley 1988), and ρ_{XOQ}^2 (Xu and O'Quigley 1999).

This line of research stems from Kent (1983)'s general idea of dependence as information gain, expressed through the Kullback–Leibler distance between models (Kullback and Leibler 1951). Kent and O'Quigley (1988) applied Kent (1983)'s approach to develop a measure known as ρ_W^2 for use in Cox PH models. ρ_W^2 is a rather complex statistic motivated by a kind of equivalence between Cox and Weibull models (the W in ρ_W^2 standing for *Weibull*). For a Cox model, ρ_W^2 involves the calculation of the expected information gain of an "equivalent" Weibull model. Because of the link with information gain (i.e., reduction in entropy), Kent and O'Quigley (1988) described ρ_W^2 as the proportion of "explained randomness" of a model, rather than explained variation.

Kent and O'Quigley (1988) showed that ρ_W^2 may be approximated by a much simpler statistic

$$\rho_{W,A}^2 = \frac{A}{1+A} \tag{4}$$

where $A = \text{var}(\mathbf{x}\boldsymbol{\beta})$. A has already appeared in this article in the context of linear regression; see (1). Kent and O'Quigley (1988) initially considered

$$\rho_{PM}^2 = \frac{A}{\sigma^2 + A} \tag{5}$$

as a possible measure of dependence. ρ_{PM}^2 can be seen as an estimate of the squared Pearson correlation between the logarithm of transformed survival time and the index $\mathbf{x}\boldsymbol{\beta}$ in a linear model with errors distributed according to a Gumbel (log-Weibull) distribution, for which the residual variance is $\sigma^2 = \pi^2/6 \simeq 1.645$. See Kent and O'Quigley (1988)'s equation (1.2) for the model underlying this interpretation, and the surrounding text for further details of ρ_{PM}^2. Because of a preference for what they considered the stronger fundamentals of information-based measures of dependence, Kent and O'Quigley (1988) rejected ρ_{PM}^2 in favor of ρ_W^2 and its approximation, $\rho_{W,A}^2$.

2.1 Explained randomness versus explained variation

For the normal-errors regression model without censoring, explained variation and explained randomness (and the resulting statistics) coincide. However, for PH models with or without censoring, the statistics are different. I conjecture that explained randomness always exceeds explained variation for PH models, which is certainly the case when considering $\rho^2_{W,A}$ (an approximation to explained randomness) versus ρ^2_{PM} (a measure of explained variation). Comparing (4) and (5) shows that if $\sigma^2 > 1$ then $\rho^2_{W,A} > \rho^2_{PM}$. For the PH model, $\sigma^2 \simeq 1.645$, so the condition on σ^2 is satisfied. For normal-errors models in standardized form, $\sigma^2 = 1$ and $\rho^2_{W,A} = \rho^2_{PM}$.

2.2 Proposed measures of explained variation

Because of its satisfying theoretical underpinning in the Kullback–Leibler distance between models, I regard ρ^2_W as the "gold standard" of explained randomness measures for survival models under the PH assumption. However, ρ^2_W is complicated to compute and does not measure explained variation. In my view, the latter property is a drawback—I suggest a simple solution below. For practical purposes, an approximation to ρ^2_W must first be considered. Of the available easy-to-calculate approximations, ρ^2_k is better than $\rho^2_{W,A}$ in two important respects. First, ρ^2_k possesses the nesting property, whereas $\rho^2_{W,A}$ does not. Second, ρ^2_k is useful in model validation. In an independent or test sample, $\rho^2_{W,A}$ depends on $\mathbf{x}\boldsymbol{\beta}$ predicted from the original training sample but not on the outcome in the test sample. Therefore, $\rho^2_{W,A}$ is useless in validation, where evaluation of the predictive ability of a predefined index is required.

The main disadvantage of ρ^2_k appears to be a mild increase (upward bias) with larger amounts of censoring (see, for example, tables I and II of O'Quigley, Xu, and Stare [2005]), whereas $\rho^2_{W,A}$ appears to be largely independent of censoring. ρ^2_W also appears to be independent of censoring, so the approximations relating ρ^2_k to ρ^2_W presumably must start to break down with increasing amounts of censoring.

To create a statistic based on ρ^2_k and resembling a measure of explained variation, while inheriting its good properties, we may reexpress ρ^2_k as follows. Suppose that we write ρ^2_k in the form $V/(1+V)$ where $V \geq 0$, so that $V = \rho^2_k \big/ \left(1-\rho^2_k\right)$. With this definition, ρ^2_k resembles $\rho^2_{W,A}$ in structure, with $V \simeq A$. By analogy with $\rho^2_{PM} = A/\left(\pi^2/6 + A\right)$, a measure with the character of explained variation in PH models may be derived as

$$R^2 = \frac{V}{\pi^2/6 + V} = \frac{\rho^2_k}{\rho^2_k + (\pi^2/6)\,(1-\rho^2_k)} \tag{6}$$

R^2 in (6) is our proposed measure of explained variation for a PH model.

Royston and Sauerbrei (2004)'s R^2_D is a transformation of the D measure. The latter is computed by ordering the estimated index $\mathbf{x}\widehat{\boldsymbol{\beta}}$, calculating the expected normal order statistics corresponding to these values, scaling the latter by dividing by a factor $\kappa = \sqrt{8/\pi} \simeq 1.60$, and performing an auxiliary regression on the scaled scores. The resulting regression coefficient is D. The conversion to R^2_D is given by

$$R^2_D = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2} \tag{7}$$

where

$$\sigma^2 = \begin{cases} 1 & \text{(lognormal model or models with a probit link)} \\ \pi^2/3 & \text{(log-logistic model or proportional odds models)} \\ \pi^2/6 & \text{(PH models)} \end{cases}$$

As may be seen by comparing (7) with (6) and (5), D^2/κ^2 plays the same mathematical role in R^2_D as do V in R^2 and A in ρ^2_{PM}. D^2/κ^2 is interpretable as an estimate of the variance of the index $\mathbf{x}\boldsymbol{\beta}$.

2.3 Other survival models

In models that do not incorporate the PH assumption for covariate effects, R^2 in (6) is not interpretable as a measure of explained variation, since the inferential basis is no longer valid. However, R^2 is still useful as a rough-and-ready index of determination in non-PH survival models. For a given covariate vector, $\mathbf{x}$, the model χ^2 statistic is often numerically roughly equal across different types of survival model, when $\boldsymbol{\beta}$ is refitted within each model framework. The magnitude of R^2 values between such models will therefore often be comparable. I give an example of this phenomenon later.

2.4 Adjusted R^2

Under H_0: $\boldsymbol{\beta} = 0$, the quantity X^2 has a positive expected value asymptotically equal to $\dim(\boldsymbol{\beta})$. Therefore R^2 will have a positive mean under H_0. In linear regression, an adjusted R^2 that has mean 0 under H_0 is available. In survival analysis, the expected value of R^2 in (6), after subtracting $\dim(\boldsymbol{\beta})$ from X^2, will be close—but not equal—to zero. The `str2ph` software has an `adjust` option to report this adjusted R^2.

A similarly motivated adjustment to R^2_D, as described by Royston and Sauerbrei (2004), is implemented in `str2d`.

3 Syntax

`str2ph` *survival_cmd varlist* [*if*] [*in*] [`, adjust validate(`*varname*`)`
`bootreps(`*#*`) calibrate` [`no`]`dots randomness` *survival_cmd_options*]

`str2d` *survival_cmd varlist* [*if*] [*in*] [`, adjust validate(`*varname*`)`
`bootreps(`*#*`)` [`no`]`dots randomness` *survival_cmd_options*]

where *survival_cmd* is an `st` survival command (`stcox`, `streg`, or `stpm`, if installed). You must `stset` your data before using `str2ph`. `str2ph` computes Royston's modification (6) of O'Quigley, Xu, and Stare (2005)'s modification of Nagelkerke (1991)'s coefficient of determination for the survival model defined by *survival_cmd varlist* [, *survival_cmd_options*]

See the `validate()` option for comments on out-of-sample prediction and assessment of R^2 in a test sample.

4 Options

`adjust` computes adjusted R^2, taking into account the dimension (i.e., number of covariates) of the model. This option may be helpful when R^2 is low and/or the model is complex, since the expected value of R^2 under the null hypothesis (that the outcome is unrelated to the covariates) is greater than zero and depends on the model dimension. Adjustment attempts to eliminate this bias in R^2 under the null hypothesis. Since R^2 calculated by out-of-sample prediction in a test sample does not require adjustment, the `validate()` option is not permitted with `adjust`.

`validate(`*varname*`)` fits the model in the subsample defined by the low value of *varname* and computes R^2 in the subsample defined by the high value of *varname*. These subsamples may be thought of as a training and a test set. *varname* must have exactly two distinct values in the estimation sample defined by *varlist* and `if` and `in`. These two values are arbitrary. *varname* may be a string variable, in which case lexicographic ordering is assumed. R^2 is computed according to the index (`xb`) predicted from the training sample (low value of *varname*) into the test sample (high value of *varname*). With `str2ph`, there is a choice between refitting the index as the only covariate in the test sample and offsetting the index there (see the `calibrate` option). With `str2d`, the index predicted on the test sample is transformed to scaled normal scores and regression on the scores is performed. The slope of this regression is Royston and Sauerbrei's D statistic. This step is required to compute D and hence R^2. Since calibration is effectively done by the regression on the scores, the `calibrate` option is not relevant to the D method and is not available with `str2d`.

`bootreps(`*#*`)`, where $\# > 0$, computes a bootstrap confidence interval for R^2 or R^2_D, using # bootstrap replications. A minimum reasonable value of # is 1000, but a better number is 5000. With $\# = 5000$, the computation may take quite some time. The default value of # is 0, meaning no confidence interval is computed by `str2ph`. With $\# = 0$ in `str2d`, an analytic estimate of the standard error (SE of R^2 is displayed, derived by the delta method from the SE of D (see Royston and Sauerbrei [2004] for more details of the SE of D).

`calibrate` (for use only with `str2ph ..., validate()`) forces the survival regression to be reestimated in the test sample on the index predicted from *varlist* in the training sample. The default is to offset the predicted index and calculate R^2 via the likelihood of that model. Regression on the index amounts to calibration of the model in the test sample and may noticeably increase the R^2 value. See also the `validate()` option.

`nodots` suppresses display of the replication dots with bootstrap confidence interval estimation. By default, a dot character is displayed after each 100 replications.

`randomness` prevents conversion of the modified Nagelkerke index of determination from explained randomness to explained variation. The reported R^2 or R^2_D is then interpretable, at least in PH models, as explained randomness.

survival_cmd_options are options of *survival_cmd*. Examples include `distribution(weibull)` for `streg`, `df(2) scale(hazard)` for `stpm`, and `strata(x1 x2)` for `stcox`.

5 Example

We will again work with the breast cancer dataset that was analyzed in detail by Sauerbrei and Royston (1999). The data are provided in `brcancer.dta` and relate to a set of 686 patients with lymph node–positive breast cancer. The outcome of interest is the recurrence-free survival (RFS) time, that is, the duration in years from entry into the study (typically, the time of diagnosis of primary breast cancer) until either death or disease recurrence, whichever occurred first. There were 299 events for this outcome and the median follow-up time was about 5 years.

Model III of Sauerbrei and Royston (1999) was a Cox proportional-hazards model for RFS that included five covariates: age (`x1`) with a fractional polynomial transformation with powers −2 and −0.5, tumor grade 2/3 (`x4a`), number of positive lymph nodes (`x5`) with the exponential transformation $\texttt{x5e} = \exp(-0.12 * \texttt{x5})$, progesterone receptors (`x6`) with a fractional polynomial transformation with power 0.5, and hormonal therapy with tamoxifen (`hormon`). Sauerbrei and Royston (1999) adjusted for `hormon` in all models, and they did not explicitly report its regression coefficient; here we have treated `hormon` simply as a covariate. Model III may be fitted in Stata as follows:

```
. use brcancer
(German breast cancer data)
. set seed 1234
. stset rectime, fail(censrec)
  (output omitted)
. fracgen x1 -2 -0.5
-> gen double x1_1 = X^-2
-> gen double x1_2 = X^-0.5
   (where: X = x1/10)
. fracgen x6 0.5
-> gen double x6_1 = X^0.5
   (where: X = (x6+1)/1000)
. str2ph stcox x1_1 x1_2 x4a x5e x6_1 hormon
R^2 (explained variation): Cox model
   Obs    Events     R^2       Boot. SE   95% conf. interval
------------------------------------------------------------
   686      299    0.289039           .          .         .
------------------------------------------------------------
```

Note the use of fracgen to transform the covariates before stcox is used (silently) to fit the Cox model within str2ph. The explained variation for the model is 0.289. The output provides no confidence interval since we did not specify the bootreps() option.

A similar format is used to compute R_D^2:

```
. str2d stcox x1_1 x1_2 x4a x5e x6_1 hormon
R^2 (explained variation - D method): Cox model
   Obs    Events     R^2       Std. err.  95% conf. interval    D      SE
---------------------------------------------------------------------------
   686      299    0.274702    0.032969   0.210028  0.338423  1.260  0.104
---------------------------------------------------------------------------
```

R_D^2 is 0.275 (SE 0.033). An estimated 95% confidence interval for R_D^2 is provided by default. Bootstrap confidence intervals are also available via the bootreps() option.

Table 1 shows values of R^2 and related statistics for a variety of models based on model III. To illustrate different variation between models and across measures, predictors were removed from model III one at a time, in the order that reduced the R^2 the least at each step, giving five models.

Table 1: Measures of explained variation and (final column) explained randomness for models for the breast cancer data

Variables in model	Deviance	R^2	95% CI	R_D^2	ρ_{PM}^2	ρ_k^2
	$-2l_{\widehat{\beta}}$	(6)	(bootstrap)[a]		(5)	(3)
x1_1 x1_2 x4a x5e x6_1 hormon	3,423.2	0.289	$0.224, 0.372$	0.275	0.271	0.401
x1_1 x1_2 x5e x6_1 hormon	3,428.2	0.281	$0.215, 0.363$	0.269	0.260	0.391
x1_1 x1_2 x5e x6_1	3,438.2	0.263	$0.197, 0.346$	0.259	0.249	0.370
x5e x6_1	3,456.9	0.230	$0.162, 0.305$	0.223	0.225	0.329
x5e	3,498.0	0.154	$0.096, 0.223$	0.141	0.137	0.231

[a] Confidence intervals were calculated by using 100 bootstrap replicates.

Most of the prognostic information is carried by x5e and x6_1, since R^2 increases only by $0.289 - 0.230 = 0.059$ on adding the variables x1_1 x1_2 x4a hormon to the model. The confidence intervals for R^2 are fairly wide. R^2, R_D^2, and ρ_{PM}^2 have similar values, with R^2 always somewhat exceeding ρ_{PM}^2 and R_D^2. The final column of table 1 shows ρ_k^2. Its values are substantially higher than those of R^2, demonstrating that explained randomness and explained variation differ markedly in PH models.

5.1 Non-PH models

Table 2 shows R^2 and R^2_D values for the covariates in model III across different types of models.

Table 2: Values of X^2 (model χ^2), R^2, and Akaike information criterion (AIC) for seven model types for breast cancer data

Model type	X^2	R^2	R^2_D	AIC
Cox	153.1	0.289	0.275	—[a]
Exponential	133.9	0.256	0.234	1,300.6
Weibull	162.6	0.305	0.288	1,252.8
Gompertz	148.4	0.281	0.267	1,285.4
`stpm` (hazard)[b]	153.2	0.289	0.275	1,217.3
Log logistic	162.6	0.305	0.253	1,234.2
`stpm` (odds)[b]	155.6	0.293	0.243	1,215.4
Lognormal	158.4	0.298	0.274	1,220.2
`stpm` (probit)[b]	155.6	0.293	0.243	1,212.2
Gamma	150.7	0.285	—[c]	1,220.3

[a] AIC is not appropriate for the Cox model in this context.

[b] Flexible parametric model with 2 degrees of freedom for the baseline distribution function, and link function as specified in parentheses.

[c] R^2_D is not available for the gamma model.

The first five rows of table 2 are for PH models; the rest are for non-PH. The R^2 values for all model types are roughly similar. Values of R^2_D are also broadly similar across models. There is close agreement between the statistics for the Cox model and the `stpm` model. This agreement is expected, since the `stpm` model is essentially a Cox model in which the baseline log cumulative hazard function is estimated as a smooth parametric (spline) function of time, whereas in the Cox model it is (or would be, if estimated explicitly) a noisy step function. Unlike the Cox model, `stpm` models are fitted by full maximum likelihood and therefore their likelihoods are comparable with those of the other parametric models.

Similarity of R^2 values across models does *not* imply that all the models fit the data equally well (see further comments in the next section). The relative fit of the models may be judged by the value of the Akaike information criterion (AIC), which is given in the final column of table 2. The AIC is the deviance (i.e., minus twice the maximized log likelihood) plus twice the model dimension. The dimension is assumed to include auxiliary parameters, such as a scale parameter, where present. The AIC for the Cox model is not reported since partial likelihood is used, and the resulting partial AIC is not comparable with AIC calculated from a full likelihood. The best-fitting model appears to be the flexible parametric model (Royston and Parmar 2002) fitted by `stpm` (Royston 2001) with a probit link function.

5.2 Comparing R^2 values: a trap for the unwary

A subtle error may be made when trying to compare R^2 values across different types of model. Higher R^2 does not necessarily mean a better fit; this relation is the case only when comparing two non-null models for which the underlying null model is identical. Consider the generalized gamma model, which contains as special cases the Weibull and lognormal models and is of course more flexible than either of them. However, as seen in table 2, the R^2 for the gamma model is *lower* than that for the Weibull or lognormal because R^2 compares the model having given covariates with that having no covariates. The gamma model fits the underlying distribution (no covariates) better than the other two models do, leaving less scope for inclusion of covariates to improve the fit than with the other models. This effect may be seen in table 3.

Table 3: Deviances and model χ^2 statistics for breast cancer–null model and model III, according to three parametric survival models

Model	Deviance		X^2
	Null model	Model III	
Gamma	1,353.0	1,202.3	150.7
Weibull	1,399.4	1,236.8	162.6
Lognormal	1,362.7	1,204.2	158.4

The gamma model has the lowest deviance for the null model and for model III, as it must have. However, the Weibull fits the distribution of log survival times in the null model much worse than the gamma and lognormal models. It therefore may have the greatest scope for improvement when covariates are added; in fact, its R^2 turns out to be the largest. Nevertheless, as judged by the AIC (see table 2), the Weibull model with covariates fits the worst of the three by some margin.

6 Validation

Model validation is an important application of R^2. Suppose that we have training and test samples, each with the same covariates recorded. A model is developed on the training data, its R^2 is estimated, and its performance is evaluated on the test data. Typically, the index $\mathbf{x}\boldsymbol{\beta}$ is estimated on the training data and this index is used, *without reestimation of* $\boldsymbol{\beta}$, to predict the outcome in the test data. The covariate values in the test data are used to compute $\mathbf{x}\boldsymbol{\beta}$ there—in Stata terminology, this is out-of-sample prediction.

Suppose that the dataset includes both training and test data, with an indicator variable, say, `test`, taking the value 0 for the training sample and 1 for the test sample. Such a variable, derived by a random split of the data, has been included in the `brcancer` dataset for exemplification.

Suppose that we wanted to evaluate the performance, according to R^2, of the prespecified model III fitted on the training data (`test==0`) in the test data (`test==1`), and get a 95% confidence interval for R^2. For the training data, we run

```
. str2ph stcox x1_1 x1_2 x4a x5e x6_1 hormon if test==0, bootreps(5000)
.................................................
R^2 (explained variation): Cox model
   Obs    Events     R^2       Boot. SE   95% conf. interval
------------------------------------------------------------
   343      148    0.320285    0.056558   0.230017  0.452113
------------------------------------------------------------
```

Now we evaluate R^2 for the index predicted on the test data:

```
. str2ph stcox x1_1 x1_2 x4a x5e x6_1 hormon, validate(test) bootreps(5000)
.................................................
R^2 (explained variation): Cox model
   Obs    Events     R^2       Boot. SE   95% conf. interval
------------------------------------------------------------
   343      151    0.216595    0.051386   0.132217  0.333522
------------------------------------------------------------
Note: model fitted at low value of test, evaluated at high value
Note: index offset from linear predictor in validation sample
```

The R^2 value on the test data is 0.217, lower than the value of 0.320 in the training sample. Since we have estimates of the SE of R^2 on both subsamples and the estimates are independent, we can also find a confidence interval for the difference between the R^2 values on the training and test data. The SE of the difference is estimated as $\sqrt{0.0565^2 + 0.0514^2} = 0.0764$ and a 95% confidence interval for the difference in R^2 is $(-0.253, 0.046)$. Since the confidence interval includes zero we might conclude that the difference in R^2 was compatible with no real difference.

We might instead wish to evaluate R^2, allowing regression on the predicted index in the test sample. This approach is a type of model calibration (e.g., Verweij and van Houwelingen [1993]). To do this, one would add the `calibrate` option to the validation run.

```
. str2ph stcox x1_1 x1_2 x4a x5e x6_1 hormon, validate(test) bootreps(5000)
> calibrate
.................................................
R^2 (explained variation): Cox model
   Obs    Events     R^2       Boot. SE   95% conf. interval
------------------------------------------------------------
   343      151    0.224467    0.050078   0.133314  0.327437
------------------------------------------------------------
Note: model fitted at low value of test, evaluated at high value
Note: calibrated on index in validation sample
```

The value of R^2 and the corresponding fit of the model have increased slightly by calibration, from 0.217 to 0.224. This result suggests that the index predicted from the subsample with `test==0` needs little calibration. The calibration slope (i.e., the shrinkage statistic $\widehat{c}$ proposed by Verweij and van Houwelingen [1993]) is 0.84 (SE 0.11) and a 95% confidence interval for c includes 1. A value of $c = 1$ means no calibration and is represented by `validate()` without the `calibrate` option. When available, $\widehat{c}$ and its SE are returned by **str2ph** in `r(c)` and `r(cse)`, respectively.

This example is merely pedagogic. A real validation exercise would have an independent test sample, not use a random subset of the original data.

7 Final comments

Once you leave the confines of the normal-errors linear model, the concept of explained variation becomes tricky. A useful summary of the situation from the Stata user's point of view is given in a frequently asked question by Nicholas Cox (2003) on the Stata web site. Cox comments that, in cases of doubt when Stata does not supply an R^2,

> There is usually something you can do for yourself: calculate the correlation between the observed response and the predicted response, and then square it.

This idea is explored for the generalized linear model by Zheng and Agresti (2000). Unfortunately, the approach does not work very naturally in survival analysis with censored observations, which is why more complicated approaches are needed.

An advantage of ρ_k^2 and R^2 is their availability with more complex models (e.g., time-varying Cox models) for which $\rho_{W,A}^2$ is undefined. In a time-varying Cox model, $\mathbf{x}\boldsymbol{\beta}$ and $\text{var}(\mathbf{x}\boldsymbol{\beta})$ depend on time. How to extend $\rho_{W,A}^2$ and R_D^2 to allow for time dependence is an open question.

Finally, more work is needed on the performance of R^2, including comparisons with R_D^2. For example, simulations to clarify the dependency of the expected value of these statistics on the amount of censoring would be helpful. My preliminary impression is that R^2 increases rather more rapidly with the amount of censoring than does R_D^2. Limited information on the relationship between D (and implicitly R_D^2) and the amount of censoring is given in tables 2 and 4 of Royston and Sauerbrei (2004).

8 Acknowledgments

I thank Babak Oskooei, Nicholas J. Cox, Willi Sauerbrei, and a referee for helpful comments on earlier versions of the manuscript.

9 References

Cox, N. J. 2003. FAQ: How can I get an R-squared value when a Stata command does not supply one? http://www.stata.com/support/faqs/stat/rsquared.html.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18: 2529–2545.

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

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

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

Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78: 691–692.

O'Quigley, J., R. Xu, and J. Stare. 2005. Explained randomness in proportional hazards models. *Statistics in Medicine* 24: 479–489.

Royston, P. 2001. Flexible parametric alternatives to the Cox model, and more. *Stata Journal* 1: 1–28.

Royston, P., and M. K. B. Parmar. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21: 2175–2197.

Royston, P., and W. Sauerbrei. 2004. A new measure of prognostic separation in survival data. *Statistics in Medicine* 23: 723–748.

Sauerbrei, W., and P. Royston. 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 162: 71–94.

Schemper, M. 1990. The explained variation in proportional hazards regression. *Biometrika* 77: 216–218.

Verweij, P. J., and H. C. van Houwelingen. 1993. Cross-validation in survival analysis. *Statistics in Medicine* 12: 2305–2314.

Xu, R., and J. O'Quigley. 1999. A measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics* 12: 83–107.

Zheng, B., and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. *Statistics in Medicine* 19: 1771–1781.

About the author

Patrick Royston is a medical statistician with 25 years of experience, with a strong interest in biostatistical methodology and in statistical computing and algorithms. He works in clinical trials and related research issues in kidney and other cancers. Currently, he is focusing on problems of model building and validation with survival data, including prognostic factors studies; on complex sample size problems in clinical trials with a survival-time endpoint; on writing a book on multivariable regression modeling; and on new trial designs.