Medical University of Vienna
Center for Medical Statistics, Informatics and Intelligent Systems
Section for Clinical Biometrics
Section Head: Prof. M. Schemper
A-1090 VIENNA, Spitalgasse 23
Phone: (+43)(1) 40400/6688
Fax: (+43)(1) 40400/6687
e-mail: biometrie@meduniwien.ac.at

Technical Report 4/2012

RELIMPCR and RELIMPLR

SAS-macros for the analysis of relative importance of prognostic factors in Cox and logistic regression

Georg Heinze* and Michael Schemper

* e-mail: georg.heinze@meduniwien.ac.at

Abstract

Two SAS macro programs are presented that compute the relative importance of covariates in the proportional hazards regression model and in the logistic regression model. The relative importance of covariates is quantified by the proportion of explained variation in the outcome (PEV) attributable to those covariates. For proportional hazards regression, the program %RELIMPCR uses the recently proposed measure V to calculate the proportion of explained variation. For the logistic model, the R^2 measure based on Pearson residuals is used by the program %RELIMPLR. Both programs are able to compute marginal and partial PEV, to compare PEV between groups of variables, and even to compare PEV between different models. They use a bootstrap resampling scheme to assess the covariance of PEV of different factors or models. Confidence limits for P-values are provided. The programs further allow to base the computation of PEV on models with shrinked or bias-corrected parameter estimates. The SAS macros are freely available at the website http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/relimp/.

Contents

1 Introduction

Statistical analysis of a prognostic factor study should involve the estimation of marginal and partial effects. Results of such an analysis are suitably presented in a table containing estimated effects, confidence limits and P-values. These standard requirements of a proper statistical analysis have been suggested by Armitage and Gehan [1] and Gehan and Walker [2].

Schemper [3] pointed out that apart from reporting marginal and partial effects estimates, it can be necessary to comment on the *relative importance* of factors in a model by computing the proportion of total variation in the outcome variable that can be explained by each factor. The concept of relative importance allows a descriptive ranking of prognostic factors according to their statistically determined importance. Using bootstrap techniques [3] it is also possible to compare the importance of different prognostic factors statistically.

This Technical Report presents the SAS [4] macro programs %RELIMPCR and %RELIMPLR that can perform these calculations for Cox and logistic regression, respectively, and are available via world wide web (see p. 10). Our programs use the most recent approaches to calculate explained variation in a Cox model [5] or a logistic model [6]. They also allow for automatic shrinkage of parameter estimates [7] and for bias-corrected estimates [8, 9] that are free from nonconvergence problems. Our programs can be run in two modes: the first mode compares the relative importance of individual prognostic factors or of groupings of such factors in the Cox or logistic model, marginally and partially ('total mode'). Run in the other mode, the programs evaluate and compare the marginal and partial relative importance of candidate factors that might be added to a given set of prognostic factors ('candidate mode'). If multiple comparisons of relative importance are done, our programs use an efficient method to correct the P-values. Table 1 compares the %RELIMPCR-macro to earlier macros (%RELIMP [3] and %SUREV [5]) for relative importance and explained variation in Cox regression that have been developed in our department. There is no predecessor to the %RELIMPLR macro.

In Section 2, we describe the measures, methods and algorithms implemented in our macros. Section 3 gives detailed information about availability, compatibility, installation, user-definable macro options, and the working modes of the macro. Section 4 shows the application of our programs by means of examples. Finally, Section 5 gives some general concluding remarks.

This Technical Report is an update of an earlier one (TR3/2001).

Table 1: Comparison of %RELIMPCR with %RELIMP and %SUREV

Property	%RELIMPCR	%RELIMP	%SUREV
Computes PEV for total model	✓	✓	✓
Computes marginal PEV automatically	✓	✓	
Computes partial PEV automatically	✓	✓	
Compares marginal PEV	✓	✓	
Compares partial PEV	✓	✓	
Compares partial PEV for candidate factors	✓		
Compares PEV for groups of factors	✓		
Confidence intervals for P-values	✓		
Adjustment for multiple testing	✓		
Computes PEV based on Firth-type estimates	✓		
Computes PEV based on shrinked estimates	✓		
Requires only SAS (without installing DLLs)*	✓	✓	✓
Uses V measure	✓		✓

* since versions 2012-04 of RELIMPCR and SUREV

2 Methods

2.1 Importance of covariates

A well-interpretable measure to quantify the importance of prognostic factors is the proportion of explained variation (PEV), i. e. the amount of variation of the outcome variable that is attributable to one or more prognostic factors, relative to the total variation of the outcome variable. While in linear models R^2 is a very popular measure of explained variation, various suggestions have been made for Cox regression (see [10, 11, 12] and others, reviewed in [13]), the most recent proposal being the measure V by Schemper and Henderson [5]. In a review of available measures of explained variation for logistic regression, Mittlböck and Schemper [6] recommend to use the R^2 measure based on squared raw residuals. From now on both measures will be denoted by R^2.

In order to define the different concepts of explained variation implemented by our programs, let us consider the following sets: let $\mathbf{S}$ denote the *standard set* of S prognostic factors considered, and let $\mathbf{I}$ denote a subset of $\mathbf{S}$. Often, but not necessarily, the set $\mathbf{I}$ will contain only one prognostic factor. Finally, let $\mathbf{C}$ denote a *candidate factor* which is not contained in the standard set $\mathbf{S}$. Partial analysis of a candidate factor $\mathbf{C}$ is adjusted for all factors in the standard set $\mathbf{S}$ but not for any other candidate factor $\mathbf{C}'$.

We now define marginal and partial PEV by $R^2_{\mathbf{I}}$ and $R^2_{\mathbf{I}|\{\mathbf{S}\setminus\mathbf{I}\}}$, respectively, where $R^2_{\mathbf{I}}$ denotes the PEV due to a model for the factor set $\mathbf{I}$ and $R^2_{\mathbf{I}|\{\mathbf{S}\setminus\mathbf{I}\}}$ is obtained by $R^2_{\mathbf{S}} - R^2_{\mathbf{S}\setminus\mathbf{I}}$. Thus, partial explained variation measures the decline in explained variation when removing the prognostic factor set $\mathbf{I}$ from a model containing the standard set $\mathbf{S}$. Similarly, partial explained variation for a candidate factor $\mathbf{C}$ is defined by $R^2_{\mathbf{C}|\mathbf{S}}$.

An analysis of relative importance of prognostic factors will comprise comparisons of PEV between different (groups of) factors, descriptively and by means of tests.

2.2 Analysis of relative importance

A descriptive comparison of the importance of prognostic factors, i. e. their relative importance, is accomplished by a table containing marginal ($\hat{R}^2_{\mathbf{I}}$) and partial ($\hat{R}^2_{\mathbf{I}|\{\mathbf{S}\setminus\mathbf{I}\}}$) PEVs for each prognostic factor or group of prognostic factors. Additionally, differences in marginal or partial PEV between disjunct factor sets $\mathbf{I}$ and $\mathbf{I}'$ could be tested statistically, dividing $\hat{D}_{\mathbf{I}\mathbf{I}'} = \hat{R}^2_{\mathbf{I}} - \hat{R}^2_{\mathbf{I}'}$ or $\hat{D}_{\mathbf{I}\mathbf{I}'} = \hat{R}^2_{\mathbf{I}|\{\mathbf{S}\setminus\mathbf{I}\}} - \hat{R}^2_{\mathbf{I}'|\{\mathbf{S}\setminus\mathbf{I}'\}}$, respectively, by their respective standard errors, and comparing this ratio to the standard normal distribution. This statistical test procedure relies on the normality of the test statistic which can be improved

by transforming the R^2 values into $M = \arcsin\sqrt{R^2}$, and redefining $D_{\mathbf{II}'} = M_{\mathbf{I}} - M_{\mathbf{I}'}$. From our experience, the angular transformation $\arcsin\sqrt{\cdot}$ better achieves approximate normality than the occasionally considered $\log[-\log(\cdot)]$ transform.

More formally, we compare the importance of two prognostic factors (or groups of prognostic factors) by the difference $D_{\mathbf{II}'} = M_{\mathbf{I}} - M_{\mathbf{I}'}$, and test the null hypothesis $H_0 : D_{\mathbf{II}'} = M_{\mathbf{I}} - M_{\mathbf{I}'} = 0$. Corresponding sample estimates are $\hat{D}_{\mathbf{II}'} = \hat{M}_{\mathbf{I}} - \hat{M}_{\mathbf{I}'}$, where $\hat{M}_{\mathbf{I}} = \arcsin\sqrt{\hat{R}^2_{\mathbf{I}}}$ is employed for a marginal, and $\hat{M}_{\mathbf{I}} = \arcsin\sqrt{\hat{R}^2_{\mathbf{S}\setminus\mathbf{I}}}$ for a partial analysis. Similarly, the partial importance of candidate factors $\mathbf{C}$ and $\mathbf{C}'$ is compared by $D_{\mathbf{CC}'} = M_{\mathbf{C}} - M_{\mathbf{C}'}$ and a statistical test of the null hypothesis $H_0 : D_{\mathbf{CC}'} = 0$ is available. Again, corresponding sample estimates are $\hat{D}_{\mathbf{CC}'} = \hat{M}_{\mathbf{C}} - \hat{M}_{\mathbf{C}'}$, where $\hat{M}_{\mathbf{C}} = \arcsin\sqrt{\hat{R}^2_{\mathbf{C}\cup\mathbf{S}}}$.

As sample estimates for the required standard errors $\hat{\sigma}_{\mathbf{II}'}$ and $\hat{\sigma}_{\mathbf{CC}'}$ are not available Schemper [3] suggested to employ the *paired bootstrap* (resampling of B samples from the original study sample) [14], p. 291. To increase efficiency, the bootstrap in our programs is operating in a balanced mode (see [14], p. 211).

Two-sided P-values for a true difference in importance of (groups of) factors $\mathbf{I}$ and $\mathbf{I}'$ are obtained from

$$\hat{P}_{\mathbf{II}'} = 2[1 - \Phi(|\hat{D}_{\mathbf{II}'}|/\hat{\sigma}_{\mathbf{II}'})]$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. The bootstrap introduces an additional element of variability into these P-values which could be removed by an infinite number for the B resamples. In agreement with the style of presentation of Monte Carlo P-values in StatXact [15] we prefer to evaluate this variability by 99% confidence limits for each of the P-values for the tests of differences in R^2. If such a confidence interval for the P-value is considered unsuitably large the program can be rerun with an increased number of resamples B.

The $(1-\alpha)\times 100\%$-confidence interval for $\hat{P}_{\mathbf{II}'}$ is computed by

$$[2\{1 - \Phi(|D_{\mathbf{II}'}|/\hat{\sigma}^*_{\mathbf{II}'})\}, 2\{1 - \Phi(|D_{\mathbf{II}'}|/\hat{\sigma}^{**}_{\mathbf{II}'})\}]$$

where

$$\hat{\sigma}^*_{\mathbf{II}'} = \hat{\sigma}_{\mathbf{II}'}\sqrt{(B-1)/\chi^2_{B-1}(1-\alpha/2)} \qquad \text{and} \qquad \hat{\sigma}^{**}_{\mathbf{II}'} = \hat{\sigma}_{\mathbf{II}'}\sqrt{(B-1)/\chi^2_{B-1}(\alpha/2)},$$

are the upper and lower confidence limits for $\hat{\sigma}_{\mathbf{II}'}$, respectively. Here, $\chi^2_\nu(\alpha)$ denotes the α-quantile of the χ^2 distribution with ν degrees of freedom. In Appendix A.1 the description of an additional SAS macro can be found that is designed for bootstrap sample size planning.

We prefer this parametric approach of testing $D_{\mathbf{II}'}$ and $D_{\mathbf{CC}'}$ over possible nonparametric tests because of its superior efficiency and also because the assumption of normality usually is realistic. Appendix A.2 gives further details on the advantage of the current approach.

2.3 Adjustment for multiple comparisons

Since in a typical prognostic factor study all prognostic factors or all groups of prognostic factors would be compared, the global significance level α will be violated. Adjustments for multiple comparisons could be done by means of the Bonferroni-Holm [16] method, which is efficient if correlations between comparisons are weak and the number of comparisons small. Since normality can be assumed, we can make use of the more efficient Tukey-Kramér [17] method, which is based on the studentized range distribution. The latter method has been implemented in our macros and can optionally be applied.

2.4 Bias and monotone likelihood

Due to special conditions in a data set finite maximum likelihood parameter estimates may not exist because the likelihood is monotone in at least one of the parameters. This breakdown of the maximum likelihood method has been termed *monotone likelihood* in Cox regression [18, 8] or *separation* in logistic regression (because of the covariate-induced separation of events and nonevents that leads to monotone likelihood) [19, 9]. Monotone likelihood particularly occurs in small samples with comparatively many estimated parameters and/or high correlation of the covariates. Even if monotone likelihood does not occur in the original sample it may affect some of the bootstrap resamples [20].

Heinze and Schemper [8, 9] showed that monotone likelihood is primarily a problem of small sample bias, and that the adaptation of a particular bias correction method due to Firth [21] guarantees finite and thus more accurate parameter estimates. Of course, their method can also be applied if separation is not present, simply to reduce the bias of parameter estimates. %RELIMPCR and %RELIMPLR both offer the option of applying Firth's bias correction.

2.5 Shrinkage

If a regression model is used to predict outcomes in future cases, then the *regression to the mean* effect implies that the future values of the response variable tend to be closer to the overall mean than might be expected from the predicted values. The extent of this

shrinkage is related to simple goodness-of-fit statistics of the original regression [7]. The predictive value of a Cox or logistic regression model can be increased if the regression coefficients related to the m covariates in the model are multiplied by the shrinkage factor $\gamma = (\chi^2_m - m)/\chi^2_m$ where χ^2_m is the model χ^2. In our macros, shrinkage can be accounted for optionally.

3 Program description

3.1 Availability and compatibility

The SAS macros %RELIMPCR, %RELIMPLR and %NBOOT (see Appendix A.1) are freely available at our WWW location http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/relimp/.

The macros have been originally developed under Windows NT and SAS 8.1, and should run under later SAS version for any platform.

3.2 Installation

3.2.1 RELIMP*CR*

The macro is distributed as a ZIP archive containing the files relimpcr.sas (the macro), example.sas (an example data set and macro call), example.lst (the output for the example macro call) and tr4_2012.pdf (this document). To install the macro, just submit the macro code relimpcr.sas in SAS. The current version, 2012-04, is written entirely in SAS/STAT and SAS/IML. (Former versions of RELIMPCR used a dynamic link library, which had to be installed.)

3.2.2 RELIMP*LR*

The ZIP archive contains the files relimplr.sas (the SAS macro), example.sas and example.lst (an example macro call and the output), and tr3_2001.pdf (this document). As the macro %RELIMPLR is written completely in SAS, it can simply be installed by copying the file 'relimplr.sas' into the folder where you store other SAS-macros.

3.3 Syntax

3.3.1 RELIMP*CR*

%RELIMPCR(
DATA=*SAS data set*,
TIME=*variable*,
CENS=*variable*,
CENSVAL=*variable*,
VARLIST=*variable list*,
CAND=*variable list*,

GROUPS=*value list*,
GNAMES=*strings*,
CLASS=*variable list*,
SHRINK=*value*,
FIRTH=*value*,
CORRECTP=*value*,
NBOOT=*value*,
ALPHA=*value*,
CLP=*value*,
PRINTCLP=*value*,
HISTOGRAM=*value*,
SEED=*value*);

3.3.2 RELIMP*LR*

%RELIMPLR(
DATA=*SAS data set*,
Y=*variable*,
VARLIST=*variable list*,
CAND=*variable list*,
GROUPS=*value list*,
GNAMES=*strings*,
CLASS=*variable list*,
SHRINK=*value*,
FIRTH=*value*,
CORRECTP=*value*,
NBOOT=*value*,
ALPHA=*value*,
CLP=*value*,
PRINTCLP=*value*,
HISTOGRAM=*value*,
SEED=*value*,
MAXIT=*value*,
MAXHS=*value*,
MAXSTEP=*value*,
EPSILON=*value*);

3.4 Description of macro options

The following macro parameters can be specified:

- DATA: names the SAS data set to be analyzed (default=_ LAST_)
- TIME: names the variable containing the survival times (only in %RELIMPCR; default=time)
- CENS: names the variable containing the status indicator (only in %RELIMPCR; default=cens)
- CENSVAL: names the code used for censored times (only in %RELIMPCR; default=0)
- Y: names the outcome variable (coded in 0 and 1) (only in %RELIMPLR; default=y)
- VARLIST: names the variable(s) to be included in the model (can be left blank if only candidate factors are to be evaluated; no default). These variables constitute the standard set **S**.
- CAND: names variable(s) that act as candidate factors (only use this option to run the programs in candidate mode; no default)
- GROUPS: specifies which variable belongs to which group (use this option to define groups of prognostic factors, do not use it otherwise). In this option, a list of group numbers should be given corresponding to the desired grouping of variables specified in VARLIST (e. g.: varlist=age sex mark1 mark2, groups=1 1 2 2; see also the example in § 4.3; by default, each prognostic factor is its own group).
- GNAMES: specifies a label for each group of variables, e. g.: gnames=environ genetic (no default)
- CLASS: names categorical variables that appear in VARLIST or CAND. These variables will automatically be dummy-coded. It is not necessary to specify two-level-factors here (no default).
- SHRINK: if set to 1, requests shrinking of parameter estimates by the factor $(\chi^2_m - m)/\chi^2_m$ where m is the number of parameters in the model and χ^2_m is the model chi-squared (default=0)

- FIRTH: if set to 1, requests Firth-type bias corrected parameter estimates (according to Heinze and Schemper [8]). If monotone likelihood is encountered, this option has to be used but may be generally preferable also in small samples or with many parameters (default=0)

- CORRECTP: indicates whether P-values should be corrected for multiple testing using Tukey's studentized range test (default=0)

- NBOOT: specifies the number of bootstrap resamples to compute standard errors (default=350). This number can be pre-planned using the SAS macro %NBOOT (see Appendix A.1). If NBOOT is set to 1, then no comparisons will be performed and the macro only computes marginal and partial PEV.

- ALPHA: the significance level (default=0.05). This option will only be used to indicate whether confidence limits for P-values include ALPHA or not.

- CLP: the confidence level for confidence limits of the P-value (default=0.99). The macro will automatically mark P-values whose confidence limits include the value ALPHA.

- PRINTCLP: if set to 1, confidence intervals for the P-values will be printed (default=0)

- HISTOGRAM: indicates whether histograms of the bootstrapped values of $D_{\mathbf{II}'}$ should be created. The macro uses PROC CAPABILITY for this option (default=0).

- SEED: specifies the random number seed used for bootstrap resampling (default=395748)

The following options are only active in the macro %RELIMPLR and apply only if Firth-type bias-corrected parameter estimates are requested by setting the option FIRTH=1:

- MAXIT: maximum number of iterations to calculate parameter estimates (default=50)

- MAXHS: maximum number of step-halvings in one iteration (default=5)

- MAXSTEP: maximum step length in one iteration (on the scale of standardized coeffients; default=5)

- EPSILON: convergence criterion for parameter estimates (default=0.0001)

3.5 Two modes of the macros

3.5.1 Candidate mode

If variables are specified in option CAND the macro will be run in 'candidate-mode', which means that marginal PEV is computed for VARLIST and CAND variables, and partial PEV is computed for CAND variables as the proportion of variation that can be explained additionally to the variation explained by all VARLIST variables together. We have applied this mode in the example of § 4.1.

3.5.2 Total mode

Not using the CAND option will run the macro in 'total mode'. In this mode, the macro will compute marginal and partial PEV for all variables specified in VARLIST and will compare marginal and partial PEV statistically. This mode will be exemplified in §§ 4.2 and 4.4.

If PEV for groups of variables should be compared, then the GROUPS option can be used to assign group numbers to the variables specified in VARLIST. Those group numbers can be mapped to group labels using the option GNAMES. The GROUPS option could be used if, e. g., a group of 'genetic' factors is to be compared with a group of 'environmental' variables in order to investigate whether genetic or environmental effects have higher impact on the survival of patients. Such an analysis is exemplified on a hypothetical study in § 4.3.

3.6 Output data sets

%RELIMPCR and %RELIMPLR produce the output data sets _pev_ and _relimp_ which may be relevant for further processing.

3.6.1 The output data set _pev_

This data set contains the following variables:

- `_sample_`: contains the number of bootstrap resample. The first line, `_sample_`=0, denotes the original data set.
- `_pevfull`: contains $R^2_{\mathbf{S}}$, the PEV for the full model (only for the original data set).

- `_asfull`: $\arcsin\sqrt{R^2_{\mathbf{S}}}$, the angular transformed PEV for the full model

If the macro runs in candidate mode, C candidate factors have been specified in CAND. The data set _ pev_ contains the following variables:

- `_pevmv1`, `_pevmv2`, ...: contain R^2_s, $s = 1, \ldots, S$; the marginal PEV for each of the k factors specified in VARLIST
- `_pevmc1`, `_pevmc2`, ...: contain R^2_c, $c = S+1, \ldots, S+C$; the marginal PEV for each of the C factors specified in CAND
- `_pevp1`, `_pevp2`, ...: contain $R^2_{\mathbf{S}\cup c}$, $c = S+1, \ldots, S+C$; the PEV for the model containing all S factors of VARLIST plus candidate factor c
- `_asmv1`, `_asmc1`, `_asp1`, ...: the angular transformed values of `_pevmv1`, `_pevmc1`, `_pevp1`, ...
- `_dm1_2`, `_dm1_3`, ..., `_dm2_3`, ...: the differences between `_asmc1` and `_asmc2`, `_asmc1` and `_asmc3`, ..., `_asmc2` and `_asmc3`, ...
- `_dp1_2`, `_dp1_3`, ..., `_dp2_3`, ...: the differences between `_asp1` and `_asp2`, `_asp1` and `_asp3`, ..., `_asp2` and `_asp3`, ...

In total mode, we have the following variables:

- `_pevm1`, `_pevm2`, ...: contain R^2_s, $s = 1, \ldots, S$; the marginal PEV for each factor specified in VARLIST
- `_pevp1`, `_pevp2`, ...: contain $R^2_{\mathbf{S}\setminus s}$, $s = 1, \ldots, S$; the PEV for the model containing all factors except the s-th ($=R^2_{\mathbf{S}} - R^2_{s|(\mathbf{S}\setminus s)}$)
- `_asm1`, `_asp1`, ...: the angular transformed values of `_pevm1`, `_pevp1`, ...
- `_dm1_2`, `_dm1_3`, ..., `_dm2_3`, ...: the differences between `_asm1` and `_asm2`, `_asm1` and `_asm3`, ..., `_asm2` and `_asm3`, ...
- `_dp1_2`, `_dp1_3`, ..., `_dp2_3`, ...: the differences between `_asp1` and `_asp2`, `_asp1` and `_asp3`, ..., `_asp2` and `_asp3`, ...

If relative importance is compared between groups of prognostic factor rather than between individual factors then the indices in these SAS variables refer to the group numbers (specified in the GROUPS option).

3.6.2 The output data set _relimp_

This data set contains all the information that is printed into the output window, and consists of one line.

Additionally to the variables `_pevfull`, `_pevmc1`, `_pevmv1`, `_pevm1`, `_pevp1`, `_asfull`, `_asmc1`, `_asmv1`, `_asm1`, `_asp1`, `dm1_2`, `dp1_2`, ..., which have the same definition as in _pev_, we have the following variables:

- `mmc1`, ...: Marginal PEV for candidate factor $S+1, \ldots, S+C$ (candidate mode)
- `mmv1`, ...: Marginal PEV for factor $1, \ldots, S$ of VARLIST
- `mpc1`, ...: Partial PEV for candidate factor $S+1, \ldots, S+C$ (candidate mode)
- `mp1`, ...: Partial PEV for factor $1, \ldots, S$ (total mode)
- `mpfull`: PEV for model with S factors of VARLIST (but without candidates)
- `sdm1_2`, ...: bootstrap estimate of the standard deviation of `dm1_2`, ...
- `sdm1_2lo`, ...: lower confidence limit for `sdm1_2`, ...
- `sdm1_2up`, ...: upper confidence limit for `sdm1_2`, ...
- `sdp1_2`, ...: bootstrap estimate of the standard deviation of `dp1_2`, ...
- `sdp1_2lo`, ...: lower confidence limit for `sdp1_2`, ...
- `sdp1_2up`, ...: upper confidence limit for `sdp1_2`, ...
- `pm1_2`, ...: P-value for comparison of marginal PEV between candidate factors/factors/groups 1 and 2, ...
- `pm1_2lo`, ...: lower confidence limit for `pm1_2`, ...
- `pm1_2up`, ...: upper confidence limit for `pm1_2`, ...
- `pp1_2`, ...: P-value for comparison of partial PEV between candidate factor/factors/groups 1 and 2, ...
- `pp1_2lo`, ...: lower confidence limit for `pm1_2`, ...
- `pp1_2up`, ...: upper confidence limit for `pm1_2`, ...

- `tm1_2`, ...: t-value for comparison of marginal PEV between candidate factors/factors/groups 1 and 2, ...
- `tm1_2lo`, ...: lower confidence limit for `tm1_2`, ...
- `tm1_2up`, ...: upper confidence limit for `tm1_2`, ...
- `tp1_2`, ...: t-value for comparison of partial PEV between candidate factors/factors/groups 1 and 2, ...
- `tp1_2lo`, ...: lower confidence limit for `tm1_2`, ...
- `tp1_2up`, ...: upper confidence limit for `tm1_2`, ...
- `hsd`: critical value from studentized range distribution

4 Examples

This section presents some example data sets analysed by RELIMPLR and RELIMPCR. Please note that in macro version 2012-04 of RELIMPCR the way output is presented was modified (adopted to SAS 9.3's default html output). However, the example output listings presented here are based on the earlier output style .

4.1 Breast cancer study 1

Latinovic et al. [22] present a comparison of the prognostic relevance of three histological grading methods in breast cancer. 292 unselected cases of node positive breast cancer were subject to surgical resection. The specimens were graded using the methods by Elston et al, by Helpap et al, by Contesso et al. using two variants, and by two variants based on the mitotic counts. Latinovic et al. investigated which of these candidate gradings best improved explanation of variation in survival times based on the known prognostic factors age, tumor stage, menopausal status and HER-2 staining intensity. While they computed Schemper and Henderson's V measure [5] of explained variation for all gradings, the question remains if the observed differences in V are only due to chance or not. Using our macro program, P-values for the comparison of explained variation can be computed. We decide to use a bootstrap sampling size of $B = 1459$ to have high precision in the estimation of the P-values (a 99% confidence interval of a P-value of 0.05 would range from 0.04 to 0.062). Our SAS data set 'breast' contains the variables 'tevent' (time from diagnosis to death or end of follow-up), 'status' (censoring status of observation), 'pt' (tumor stage), 'agegrp' (age group; coded as 1 if age>60 and 0 else), 'hercep' (HER-2 staining intensity, dichotomized to 1 and 0), 'menopaus' (menopausal status; 1 if patient is post-menopausal, 0 else); 'elston', 'contess1', 'contess2', 'helpap', 'mithpf1', and 'mithpf2' denote the different grading methods. The macro %RELIMPCR is called by submitting

```
%relimpcr(data=breast, time=tevent, cens=status,
          varlist=pt agegrp hercep menopaus,
          cand=elston contess1 contess2 helpap mithpf1 mithpf2,
          class=elston contess1 hercep mithpf1 mithpf2,
          nboot=1459, seed=573948, printclp=1);
```

The variables elston, contess1, mithpf1, mithpf2 and hercep denote categorical factors with more than two levels, therefore we have to specify them in the class option. Fig. 1, 2 and 3 show the output of the SAS macro %RELIMPCR. First we learn that the proportion

Figure 1: Output of %RELIMPCR for breast cancer study 1.

```
Proportion of Explained Variation (PEV)
=======================================

PEV          Marginal   Partial
========     ========   =======

pt             2.05%
agegrp         0.15%
hercep         1.86%
menopaus       0.20%
--------------------
Model          4.58%

Candidates:
===========

elston         5.51%     4.36%
contess1       2.91%     1.94%
contess2       4.44%     3.42%
helpap         2.76%     2.28%
mithpf1        3.04%     2.30%
mithpf2        3.70%     2.99%
```

of variation in survival times that can be explained by prognostic factors is very low. It is only 4.6% in this example. In studies of survival this value typically ranges between 10% and 35%. The proportion of 4.6% can be increased by 4.4% if the grading method by Elston is applied. All other grading methods lead to lower improvements of explained variation. The Elston method is significantly better than the Contesso 1 and mitotic count variant 1 methods, and the Helpap method. Comparing Contesso 1 and 2, we see that the second variant is relatively more important for predicting survival than the first. The same is true for the two mitotic count methods. This may be due to the microscope used: both Contesso 2 and mitotic 2 use a microscope with a high power field (HPF) of 0.238 mm^2, while Contesso 1 and mitotic 1 use an HPF of 0.345 mm^2. Fig. 3 shows the 99% confidence intervals for the P-values.

While these results are quite interesting, one may argue that since we were 'looking for' significant differences, the P-values should be adjusted for multiple comparisons. Applying the Tukey-Kramér method by specifying the option 'correctp=1' in the macro

Figure 2: Output of %RELIMPCR for breast cancer study 1 (cont.).

```
All comparisons based on 1459 bootstrap replicates.


Comparison (p-values): marginal PEV
===================================

         elston   contess1 contess2 helpap   mithpf1  mithpf2

elston    .       0.0476*  0.5259   0.1018   0.0869   0.2155
contess1  .        .       0.0597*  0.8847   0.9143   0.5318
contess2  .        .        .       0.2446   0.3301   0.5892
helpap    .        .        .        .       0.8403   0.5304
mithpf1   .        .        .        .        .       0.3705
mithpf2   .        .        .        .        .        .


Comparison (p-values): partial PEV
==================================

         elston   contess1 contess2 helpap   mithpf1  mithpf2

elston    .       0.0053   0.3822   0.0467*  0.0265   0.1535
contess1  .        .       0.0018   0.4961   0.5691   0.1289
contess2  .        .        .       0.1389   0.1637   0.5915
helpap    .        .        .        .       0.9873   0.3896
mithpf1   .        .        .        .        .       0.1011
mithpf2   .        .        .        .        .        .

*: 99%-C.L. for p-value contains the critical value 0.05.
```

Figure 3: Output of %RELIMPCR for breast cancer study 1 (cont.).

```
99%-confidence intervals for P-values
=====================================

                          Marginal PEV          Partial PEV
                        =================     =================

elston    contess1      [0.0380 , 0.0593]     [0.0035 , 0.0080]
elston    contess2      [0.5064 , 0.5459]     [0.3600 , 0.4054]
elston    helpap        [0.0865 , 0.1193]     [0.0372 , 0.0583]
elston    mithpf1       [0.0729 , 0.1031]     [0.0201 , 0.0347]
elston    mithpf2       [0.1945 , 0.2383]     [0.1349 , 0.1742]
contess1  contess2      [0.0486 , 0.0730]     [0.0011 , 0.0029]
contess1  helpap        [0.8792 , 0.8902]     [0.4759 , 0.5170]
contess1  mithpf1       [0.9102 , 0.9184]     [0.5509 , 0.5878]
contess1  mithpf2       [0.5124 , 0.5516]     [0.1117 , 0.1483]
contess2  helpap        [0.2229 , 0.2679]     [0.1211 , 0.1589]
contess2  mithpf1       [0.3077 , 0.3538]     [0.1446 , 0.1849]
contess2  mithpf2       [0.5716 , 0.6072]     [0.5740 , 0.6093]
helpap    mithpf1       [0.8328 , 0.8479]     [0.9867 , 0.9879]
helpap    mithpf2       [0.5110 , 0.5502]     [0.3675 , 0.4127]
mithpf1   mithpf2       [0.3482 , 0.3938]     [0.0859 , 0.1185]
```

call, none of the comparisons reaches significance.

The results of the comparisons are depending on the normal distribution of the differences of bootstrapped angular transformed R^2 values. The assumption can be verified graphically by plotting a histogram for each comparison. The graphical check can be obtained automatically by specifying the macro option 'histogram=1'.

4.2 Breast cancer study 2

Lösch et al. [23] present data from a breast cancer study. Survival times of 100 patients were recorded (74 of them censored) and also values of four potential risk factors: tumour stage ('T'), nodal status ('N'), histological grading ('G') and Cathepsin D immunoreactivity ('CD'). For analysis these factors were dichotomised to levels of 0 and 1 (unfavourable). The survival and censoring times are stored in variable 'months', with the censoring indicator 'cens', 0 indicating a censored survival time. The variable 'G' has an infinite maximum likelihood estimate and therefore also an infinite risk ratio estimate. Heinze and Schemper [8] present a re-analysis of this data set where they adopt a penalized maximum likelihood technique by Firth [21] to Cox regression in order to achieve finite estimates. Their technique can be applied with the %RELIMPCR-macro by setting option firth=1. Suppose the data is stored in a SAS data set called 'breast'. To evaluate the relative importance of the prognostic factors in the study, we call the macro by submitting

```
%relimpcr(data=breast, time=months, cens=cens, varlist=t n g cd, nboot=350,
          histogram=1, firth=1);
```

There are no factors in the data set that have more than two levels, therefore we do not need to specify any variables in the class option. Computation of Firth-type parameter estimates is requested, and histograms of the bootstrapped transformed R^2 values are plotted in order to graphically verify the assumption of normality. Fig. 4 shows the output of the macro. We learn that the proportion of variation explained by the multiple Cox model is about 24% of the total variation in survival. Differences in the importance of the four prognostic factors appear to be by chance only.

Figure 4: Output of %RELIMPCR for breast cancer study 2.

```
Proportion of Explained Variation (PEV)
=======================================

Parameter estimates based on penalized maximum likelihood.

PEV         Marginal   Partial
========    ========   =======

t             12.81%     4.30%
n              8.42%     2.56%
g              8.05%     1.46%
cd             7.99%     1.14%
-------------------------------
Model                   23.64%


All comparisons based on 350 bootstrap replicates.

Comparison (p-values): marginal PEV
===================================

         t        n        g        cd

t         .       0.5361   0.3446   0.4625
n         .        .       0.9437   0.9443
g         .        .        .       0.9905
cd        .        .        .        .


Comparison (p-values): partial PEV
==================================

         t        n        g        cd

t         .       0.7685   0.5020   0.5466
n         .        .       0.8047   0.7576
g         .        .        .       0.9182
cd        .        .        .        .
```

4.3 Hypothetical genetic/environmental factors study

Our third example is a hypothetical study the aim of which is to evaluate the influence of two genetic polymorphisms (poly1 and poly2) and three (dichotomous) common environmental factors (en1, en2 and en3) on the survival of 300 patients. Each genetic polymorphism can assume three different levels: AA, Aa or aa. Poly1 and Poly2 have genotype freqencies 148, 124, 28 and 113, 142, 45, respectively. 35%, 54% and 64% of the patients are exposed to the environmental risk factors ex1, ex2 and ex3, respectivley. The survival time and censoring indicator are stored in variables t and cens. For simplicity, we neither assume any interactions between the genetic factors or between environmental factors nor any genetic-environment interactions.

The Cox model, fitted by calling the PROC PHREG of SAS, shows highly significant effects of all factors involved (see Fig. 5). In this figure, the dummy variables po11 and po12 correspond to the 3-level factor poly1, and the dummies po21 and po22 correspond to poly2. The dummy codings of both variables use the AA genotype as reference group.

Now the question remains whether the genetic or the environmental factors contribute more to the explanation of variation of individual survival times. This question can be answered by applying the %RELIMPCR macro in 'group mode', assigning the polymorphisms to one 'Genetic' group and the environmental variables to another 'Environ'(mental) group:

```
%relimpcr(data=genenv, time=t, cens=cens, varlist=poly1 poly2 en1 en2 en3,
          groups=1 1 2 2 2, gnames=Genetic Environ, class=poly1 poly2,
          nboot=1459, printclp=1);
```

Again, we decide to use a bootstrap sampling size of $B = 1459$ to have high precision in the estimation of the P-values (a 99% confidence interval of a P-value of 0.05 would range from 0.04 to 0.062). The macro produces a table (Fig. 6) confirming that the environmental factors contribute 17% to the variation in survival additionally to the genetic factors, and that the genetic factors contribute 9.7% additionally to the environmental factors. Furthermore, we learn that the difference in partial PEV is statistically significant at the 5% level.

Note that in this example the marginal R^2's are lower than the respective partial R^2's for both factor sets. This may appear unusual but happens if, as in our hypothetical study, the correlation of the linear predictors for the genetic and the environmental factors is negative.

Figure 5: Cox model for the hypothetical genetic/environmental factors study

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
po11	1	0.53423	0.14009	14.5422	0.0001	1.706
po12	1	0.88096	0.22156	15.8099	<.0001	2.413
po21	1	0.70609	0.14338	24.2510	<.0001	2.026
po22	1	1.23530	0.19750	39.1209	<.0001	3.439
en1	1	1.03298	0.13778	56.2067	<.0001	2.809
en2	1	0.87525	0.13146	44.3256	<.0001	2.399
en3	1	0.53334	0.13793	14.9527	0.0001	1.705

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
poly1	22.8940	2	<.0001
poly2	44.1965	2	<.0001

Figure 6: Output of %RELIMPCR for the hypothetical genetic/environmental factors study

```
 Proportion of Explained Variation (PEV)
========================================

PEV         Marginal   Partial    Variables
========    ========   =======    =========

Genetic       7.15%     9.70%    poly1 poly2
Environ      14.47%    17.01%    en1 en2 en3
------------------------------
Model        24.17%

All comparisons based on 1459 bootstrap replicates.

Comparison (p-values): marginal PEV
===================================

         Genetic  Environ

Genetic    .       0.0285
Environ    .        .

Comparison (p-values): partial PEV
==================================

         Genetic  Environ

Genetic    .       0.0285
Environ    .        .

*: 99%-C.L. for p-value contains the critical value 0.05.

99%-confidence intervals for P-values
=====================================

                       Marginal PEV           Partial PEV
                     ================     ================

Genetic  Environ      [0.0217 , 0.0370]     [0.0217 , 0.0370]
```

Figure 7: Odds ratio estimates from PROC LOGISTIC for the low birth weight study

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.955	0.887	1.027
LWD	2.321	1.048	5.139
RACE 1 vs 3	0.442	0.185	1.059
RACE 2 vs 3	1.294	0.464	3.615
SMOKE	2.242	1.015	4.953
PTD	3.603	1.456	8.912
HT	4.201	1.179	14.966
UI	1.930	0.773	4.817

4.4 Low birth weight study

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Variables which were thought to be of importance were age, weight of the mother at her last menstrual period (LWD, dichotomized at 110 pounds), race (1=white, 2=black, 3=other), history of premature labor (PTD; 0=none, 1=at least one), history of hypertension (HT; 1=Yes, 0=No), presence of uterine irritability (UI; 1=Yes, 0=No), and smoking status during pregnacy (SMOKE; 1=Yes, 0=No). Data were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986, are given in Hosmer and Lemeshow [24] and can be downloaded from `http://www-unix.oit.umass.edu/~statdata/data/`. Calling PROC LOGISTIC we learn from the table of odds ratio estimates and confidence limits (Fig. 7) that those variables have more or less significant effects on birth weight. The model is significant with a model χ^2 of 37.1 at 8 degrees of freedom ($p < 0.0001$).

Using the macro %RELIMPLR, we can compare the importance of those variables for explaining variation in the outcome. A naive call of the macro with 350 bootstrap replications will yield some warning messages in the SAS log file saying that maximum likelihood parameter estimates of some bootstrap resamples may not exist. It is safer to use Firth's separation-proof bias correction method which has been included as option in the macro although there is no separation in the original data set. This is done by submitting the macro call

```
%relimplr(data=lowbwt, y=low, varlist=age lwd race smoke ptd ht ui,
          class=race, nboot=350, firth=1);
```

Fig. 8 shows the results of the analysis of relative importance for this study. We learn that the proportion of variation in individual birth weight that can be explained by the prognostic factors is rather low, only 18% (this should not be confused with the *goodness-of-fit*, see e. g. [25]). Furthermore, it is not possible to confirm statistically that the importance of premature labor is higher than the importance of any other prognostic factor in a model for birth weight.

Figure 8: Output of %RELIMPLR for the low birth weight study

```
Proportion of Explained Variation (PEV)
=======================================
Parameter estimates based on penalized maximum likelihood.

PEV         Marginal   Partial
========    ========   =======
age            1.29%     0.10%
lwd            4.69%     1.72%
race           2.65%     2.64%
smoke          2.61%     1.61%
ptd            7.28%     3.96%
ht             2.32%     2.45%
ui             2.86%     0.63%
------------------------------
Model         17.94%

All comparisons based on 350 bootstrap replicates.

Comparison (p-values): marginal PEV
===================================
         age      lwd      race     smoke    ptd      ht       ui
age       .       0.2894   0.5576   0.6256   0.1430   0.6879   0.5622
lwd       .        .       0.5604   0.6105   0.6073   0.5536   0.6485
race      .        .        .       0.9907   0.3107   0.9137   0.9481
smoke     .        .        .        .       0.2479   0.9324   0.9442
ptd       .        .        .        .        .       0.2816   0.3237
ht        .        .        .        .        .        .       0.8824
ui        .        .        .        .        .        .        .

Comparison (p-values): partial PEV
==================================
         age      lwd      race     smoke    ptd      ht       ui
age       .       0.5156   0.3520   0.4940   0.1955   0.3316   0.7915
lwd       .        .       0.7694   0.9722   0.5409   0.8077   0.6961
race      .        .        .       0.6936   0.7247   0.9522   0.4653
smoke     .        .        .        .       0.4764   0.7677   0.6975
ptd       .        .        .        .        .       0.6878   0.3406
ht        .        .        .        .        .        .       0.4875
ui        .        .        .        .        .        .        .
```

5 Concluding remarks

The relative importance of prognostic factors in regression models can neither be addressed by P-values nor by estimates of regression parameters that depend on the scales of measurement of prognostic factors. It can be described by explained variation and standardized regression coefficients. Both lead to similar results for continuous or dichotomous factors. However, if qualitative and quantitative factors or groups of factors are compared then explained variation is better, as the effect of any factor or group of factors can be characterized by a single number; this is not possible with regression coefficients. Therefore only explained variation has been considered here.

Evaluation of explained variation requires that a properly specified regression model has been obtained. In many applications the mechanisms influencing outcome may be too complex to be summarized by a model containing main effects only. In such cases it is possible to estimate explained variation for interaction effects, non-linear effects or time-dependent effects. If such effects become dominant, for example with qualitative or strong quantitative interactions, then, as with conventional analysis, evaluation of explained variation by subgroups may be indicated. Occasionally, in particular with highly correlated and/or interacting factors, it makes sense to evaluate explained variation for groups of such factors or for an aggregated factor, constructed from others.

The programs presented produce estimates of marginal and of partial explained variation for each of a set of prognostic factors considered. If such factors are correlated then marginal and partial explained variation will differ in the same way as marginal and partial parameter estimates or P-values will differ. There is no general preference towards the marginal or partial importance of factors, the former giving the apparent importance, without adjustments, the latter artificially keeping constant the effect of other factors. Depending on the reasons for associations among prognostic factors either marginal or partial importance may be more relevant in certain applications.

Investigation of the relative importance of prognostic factors is primarily of interest to samples, representative for a well-defined target population. In designed experiments the relative importance of a factor may depend on the particular choice of its levels.

The concepts of relative importance and of explained variation have a long tradition with linear, normal error models. Though similarly useful and intuitive for logistic and Cox regression, applicability has been limited by missing powerful computer implementations. These have now been presented and made available.

References

[1] P. Armitage and E. A. Gehan. Statistical methods for the identification and use of prognostic factors. *International Journal of Cancer*, 13:16–36, 1974.

[2] E. A. Gehan and M. D. Walker. Prognostic factors for patients with brain tumors. *National Cancer Institute Monographs*, 46:189–195, 1978.

[3] M. Schemper. The relative importance of prognostic factors in studies of survival. *Statistics in Medicine*, 12:2377–2382, 1993.

[4] SAS Institute. *SAS/STAT User's Guide, Version 8*. SAS Institute Inc., Cary, NC, 1999.

[5] M. Schemper and R. Henderson. Predicitve accuracy and explained variation in Cox regression. *Biometrics*, 56:249–255, 2000.

[6] M. Mittlböck and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15:1987–1997, 1996.

[7] J. B. Copas. Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, 6:167–183, 1997.

[8] G. Heinze and M. Schemper. A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 57(1):114–119, 2001.

[9] G. Heinze and M. Schemper. A solution for the problem of separation in logistic regression. *Statistics in Medicine*, 21:2409-2419, 2002.

[10] J.T. Kent and J. O'Quigley. Measures of dependence for censored survival data. *Biometrika*, 75:523–534, 1988.

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

[12] M. Schemper. The explained variation in proportional hazards regression. *Biometrika*, 77:216–218, 1990 (Correction: *Biometrika*, 81:631, 1994).

[13] M. Schemper and J. Stare. Explained variation in survival analysis. *Statistics in Medicine*, 15:1999–2012, 1996.

[14] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer, New York, 1995.

[15] C. Mehta and N. Patel. *StatXact4 for Windows User Manual*. Cytel Software Corporation, Cambridge, MA, 2000.

[16] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[17] C. Y. Kramer. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 12:307–310, 1956.

[18] M. C. Bryson and M. E. Johnson. The incidence of monotone likelihood in the Cox model. *Technometrics*, 23(4):381–383, 1981.

[19] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[20] T. M. Loughin. On the bootstrap and monotone likelihood in the Cox proportional hazards regression model. *Lifetime Data Analysis*, 4:393–403, 1998.

[21] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[22] L. Latinovic, G. Heinze, P. Birner, H. Samonigg, H. Hausmaninger, E. Kubista, W. Kwasny, M. Gnant, R. Jakesz, and G. Oberhuber for the Austrian Breast and Colorectal Cancer Study Group. Prognostic relevance of three histological grading methods in breast cancer. *International Journal of Oncology*, 19:1271–1277, 2001.

[23] A. Lösch, C. Tempfer, P. Kohlberger, E. A. Joura, M. Denk, B. Zajic, G. Breitenecker, and C. Kainz. Prognostic value of cathepsin D expression and association with histomorphological subtypes in breast cancer. *British Journal of Cancer*, 78:205–209, 1998.

[24] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989.

[25] M. Mittlböck and H. Heinzl. A note on R^2 measures for Poisson and logistic regression models when both models are applicable.

Journal of Clinical Epidemiology, 54:99–103, 2001.

[26] R. A. Thisted. *Elements of Statistical Computing*. Chapman and Hall, New York, 1988.

A Appendix

A.1 Bootstrap sample size considerations

The width of confidence intervals around P-values depends on the number of bootstrap resamples which are used in an analysis. This number can be pre-planned in order to achieve arbitrarily accurate bootstrap P-values. The methodology of medical study sample size calculations for confidence intervals around means can be applied. There, an *expected mean* has to be specified. In our case, when deciding on the appropriate number of bootstrap resamples, we need to replace the expected mean by the so-called *target P-value*. Usually, the target P-value will be chosen to be the significance level because we want to pre-plan the precision of P-values in that region. The other parameters needed for sample size planning are the maximally tolerable distance between the target P-value and the lower confidence bound, and the confidence level of that distance.

A.1.1 Method

Let P and L denote a target P-value and a desired lower limit of the $(1-\alpha)$ 100% confidence interval around the P-value, respectively. Then the values of the respective test statistics T_P and T_L corresponding to these P-values are

$$T_P = \Phi^{-1}(1 - P/2) = |\hat{D}_{\mathbf{II}'}|/\hat{\sigma}_{\mathbf{II}'}$$

and

$$T_L = \Phi^{-1}(1 - L/2) = |\hat{D}_{\mathbf{II}'}|/\hat{\sigma}^*_{\mathbf{II}'}$$

with $D_{\mathbf{II}'}$, $\hat{\sigma}_{\mathbf{II}'}$ and $\hat{\sigma}^*_{\mathbf{II}'}$ denoting the difference, the standard deviation of the difference and the lower $(1-\alpha)100\%$ confidence limit of the standard deviation of the difference, respectively, of the angular transformed marginal or partial R^2 values for the prognostic factor sets $\mathbf{I}$ and $\mathbf{I}'$; and as usual $\Phi^{-1}(\alpha)$ denotes the α-quantile of the standard normal distribution. The estimates $\hat{\sigma}_{\mathbf{II}'}$ and $\hat{\sigma}^*_{\mathbf{II}'}$ are obtained by the bootstrap. Since

$$\hat{\sigma}^*_{\mathbf{II}'} = \hat{\sigma}_{\mathbf{II}'}\sqrt{(B-1)/\chi^2_{B-1}(1-\alpha/2)}$$

and therefore

$$\hat{\sigma}^*_{\mathbf{II}'}/\hat{\sigma}_{\mathbf{II}'} = \sqrt{(B-1)/\chi^2_{B-1}(1-\alpha/2)}$$

and

$$\frac{\Phi^{-1}(1-P/2)}{\Phi^{-1}(1-L/2)} = \frac{T_P}{T_L} = \frac{|D_{\mathbf{II}'}|/\hat{\sigma}_{\mathbf{II}'}}{|D_{\mathbf{II}'}|/\hat{\sigma}^*_{\mathbf{II}'}} = \frac{\hat{\sigma}^*_{\mathbf{II}'}}{\hat{\sigma}_{\mathbf{II}'}}$$

we can compute B as solution to

$$\frac{B-1}{\chi^2_{B-1}(1-\alpha/2)} = \left\{\frac{\Phi^{-1}(1-P/2)}{\Phi^{-1}(1-L/2)}\right\}^2$$

The estimation could be carried out by finding the root of the target function

$$f(B, L, P, \alpha) = \frac{\chi^2_{B-1}(1-\alpha/2)}{(B-1)} - \left\{\frac{\Phi^{-1}(1-L/2)}{\Phi^{-1}(1-P/2)}\right\}^2 \tag{1}$$

using bisection (see Thisted [26]). The estimation process can be stopped, e. g., if two consecutive iterations yield the same value for the lowest integer greater or equal B. The target function (1) can also be used to find values of L or α given the others.

A.1.2 Program description

We have written the SAS [4] macro program %NBOOT for bootstrap sample size calculation. If a *target P-value* is given, it computes one of the following parameters given the other two:

- the number of bootstrap resamples
- the confidence level
- the distance between the confidence limit and the target P-value

The macro also computes the efficiency of the parametric P-value $\hat{P}$ compared to the nonparametric P-value $\tilde{P}$ (see Appendix A.2) and is available at our website (see p. 10). The syntax of the macro is as follows:

%NBOOT(P_TRUE=*value*, P_LIMIT=*value*, CONLEVP=*value*, NBOOT=*value*, MAXITER=*value*, SIDE=*string*);

- P_TARGET: specifies P, the target P-value (default=0.04)
- P_LIMIT: specifies L, the desired lower or upper limit of the confidence interval around the target P-value
- SIDE: defines whether the lower (SIDE=lower) or upper (SIDE=upper) limit is to be calculated. This option is only used if P_LIMIT has been left blank. The default value is SIDE=lower.
- CONLEVP: specifies the confidence level

- NBOOT: specifies the number of bootstrap resamples
- MAXITER: specifies the maximum number of iterations (default=50)

Among the options NBOOT, P_LIMIT and CONLEVP one has to remain unspecified. That one will be calculated by the macro.

A.1.3 Example

Suppose we want to estimate the bootstrap sample size needed to achieve a lower 99% confidence limit of 0.04 at an expected P-value of 0.05. The macro is called as follows:

```
%nboot(p_target=0.05,  p_limit=0.04, conlevp=0.99);
```

The iteration history is printed into the log file:

```
iter=1 opt=0.4985547461 n=50
iter=2 opt=-0.001264171 n=1495.9439253
iter=3 opt=-0.001142562 n=1492.2867598
iter=4 opt=0.0000226505 n=1457.9264805
iter=5 opt=-3.986371E-7 n=1458.5944074
iter=6 opt=-1.36935E-10 n=1458.5828555
```

In the above table, the variable 'opt' denotes the target function (1). After the root of the target function has been found, the macro produces the following table:

Target p-value	Limit of C.I.	Confidence level	Bootstrap resamples needed	Efficiency of normal/nonparametric
0.05	0.04	99.00%	1458.58	2.16078

# Iterations	Target function
6	-1.3693E-10

We learn that about 1459 bootstrap resamples are needed to achieve the desired precision in P-values. If the length of the confidence interval is doubled (lower precision of P-value), the bootstrap sample decreases to 350. If it is halved (higher precision), it increases to 6401. The parameter 'Efficiency of normal/nonparametric' is discussed in the following section of the Appendix.

A.2 Relative efficiency of parametric to nonparametric bootstrap

Instead of assuming a normal distribution, one could also estimate a nonparametric two-sided P-value $\tilde{P}_{\mathbf{II}'}$ by

$$\tilde{P}_{\mathbf{II}'} = 2 \min \left\{ \sum_{b=1}^{B} I(D^b_{\mathbf{II}'} \geq D_{\mathbf{II}'}), \sum_{b=1}^{B} I(D^b_{\mathbf{II}'} \leq D_{\mathbf{II}'}) \right\} / B,$$

with $I(\cdot)$ denoting the indicator function. The standard deviation of $\tilde{P}_{\mathbf{II}'}$ can be estimated by

$$\tilde{\sigma}_{\mathbf{II}'} = \sqrt{\tilde{P}_{\mathbf{II}'}(1 - \tilde{P}_{\mathbf{II}'})/(B-1)}$$

A $(1-\alpha)100\%$ confidence interval for $\tilde{P}$ is given by

$$[\tilde{P}_{\mathbf{II}'} + \tilde{\sigma}_{\mathbf{II}'}\Phi^{-1}(\alpha/2), \tilde{P}_{\mathbf{II}'} + \tilde{\sigma}_{\mathbf{II}'}\Phi^{-1}(1-\alpha/2)]$$

with $\Phi^{-1}(\alpha)$ denoting the α-quantile of the standard normal distribution. Assuming an expected P-value P and a desired lower or upper limit of the $(1-\alpha)100\%$ confidence interval L, the required bootstrap sample size for the nonparametric approach can be calculated as

$$B = \tilde{P}_{\mathbf{II}'}(1 - \tilde{P}_{\mathbf{II}'}) \left\{ \Phi^{-1}(\alpha/2) \right\}^2 / (P - L)^2 + 1$$

As we can assume the normal approximation based on the angular transformation R^2 described above to work well, both approaches will yield similar P-value estimates. Regarding the length of a confidence interval about P, the normal approximation using the angular transformation $\hat{P}_{\mathbf{II}'}$ should be preferred. This may be demonstrated by the example of Appendix A.1: given a target P-value of 0.05 and a desirable distance to a lower 99% confidence bound of not more than 0.01, a bootstrap sample size of 1459 is needed. If using $\tilde{P}$ instead of $\hat{P}$, a sample size 2.16 times higher must be used (see the table entry 'Efficiency of normal/nonparametric' in the output of the example of Appendix A.1). The efficiency gain of using $\hat{P}$ compared to $\tilde{P}$ decreases with increasing desired precision of the P-values, i. e., with increasing bootstrap sample size.