Macro Risk_Est_PH_reg_prop_score Version 1.0 User's Guide

Michael Crager
14 April 2022

Statistical Methods

Suppose we want to estimate the risk of an event occurring by a specified time using a Cox proportional hazards regression model with propensity score-based weights estimated from the same data set using a logistic regression model. Denote the regression parameter estimate for the Cox model with covariate vector $\mathbf{z}^{(\beta)}$ by $\hat{\boldsymbol{\beta}}$. The vector $\mathbf{z}^{(\beta)}$ may be time-invariant or an externally time-dependent $\mathbf{z}^{(\beta)} = \mathbf{z}^{(\beta)}(t)$, meaning that the time-dependence is fixed rather than deriving from repeated assessments of a subject's covariate values over time.

In conventional inverse probability of treatment weighting (IPTW), the weight for subject $i = 1, 2, \ldots, n$ is estimated from a logistic regression as the inverse of the estimated probability of (propensity for) the treatment received by the subject. A multinomial logistic regression with generalized logit link function can be used to accommodate $K \geq 2$ treatments. Using treatment K as the reference, let $\hat{\boldsymbol{\alpha}} = \left(\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}_2^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\alpha}}_{K-1}^{\mathrm{T}}\right)^{\mathrm{T}}$ be the vector containing all the maximum likelihood logistic regression parameter estimators, $\hat{\boldsymbol{\alpha}}_k$ being the vector of regression parameter estimators for treatment k. The probability of receiving treatment k when the covariate value is $\mathbf{z}^{(\alpha)}$ is estimated consistently by

$$\hat{p}_k(\mathbf{z}^{(\alpha)}) = \begin{cases} \dfrac{\exp\left(\hat{\boldsymbol{\alpha}}_k^{\mathrm{T}} \mathbf{z}^{(\alpha)}\right)}{1 + \sum_{j=1}^{K-1} \exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}} \mathbf{z}^{(\alpha)}\right)} & \text{for } k \leq K-1 \\ \dfrac{1}{1 + \sum_{j=1}^{K-1} \exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}} \mathbf{z}^{(\alpha)}\right)} & \text{for } k = K. \end{cases}$$

Let k_i represent the treatment received by study subject i and $\mathbf{z}_i^{(\alpha)}$ be the subject's covariate vector value (including the intercept). The estimated weight for subject i is then

$$\hat{w}_i = \frac{1}{\hat{p}_{k_i}(\mathbf{z}_i^{(\alpha)})}.$$

Use of these weights allows the practitioner to estimate the population average treatment effect (ATE) and the average event risk. These estimates will be approximately unbiased under the assumption that there are no unmeasured confounders (Rosenbaum and Rubin 1983). Stabilization of the IPTW weights by multiplying them by the proportion π_{k_i} of study subjects who received the same treatment that subject *i* received is often recommended in order to avoid very high weights and resulting inflation of the variance of model parameter estimates (Cole and Hernán 2003, Austin and Stuart 2015). The stabilized weight used in the Cox regression is $\hat{\omega}_i = \pi_{k_i}\hat{w}_i$ If the data come from a stratified cohort sampling design with sampling weight s_i (the inverse of the stratum-specific sampling ratio) for subject *i*, then the weight used in the Cox regression is $\hat{\omega}_i = s_i\pi_{k_i}\hat{w}_i$. We will use this form of the weights in the following development. If cohort sampling is not used, we set $s_i \equiv 1$. If the propensity weights are not stabilized, we set $\pi_{k_i} \equiv 1$.

For subjects with $k_i \le K-1,$ the gradient of $\hat{\omega}_i$ with respect to $\hat{\boldsymbol{\alpha}}$ is

$$\begin{aligned}
\nabla_{\hat{\boldsymbol{\alpha}}}\hat{\omega}_i &= s_i\pi_{k_i}\nabla_{\hat{\boldsymbol{\alpha}}}\frac{1+\sum_{j=1}^{K-1}\exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)} \\
&= s_i\pi_{k_i}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)\nabla_{\hat{\boldsymbol{\alpha}}}\sum_{j=1}^{K-1}\exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)-\left\{1+\sum_{j=1}^{K-1}\exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)\right\}\nabla_{\hat{\boldsymbol{\alpha}}}\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(2\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}^{(\alpha)}\right)} \\
&= s_i\pi_{k_i}\left[\left(\mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)},\mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_2^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)},\ldots,\mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_{K-1}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}\right)^{\mathrm{T}}\right. \\
&\qquad\left.-\frac{1+\sum_{j=1}^{K-1}\exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}\left(I_{\{k_i=1\}}\mathbf{z}_i^{(\alpha)\mathrm{T}},I_{\{k_i=2\}}\mathbf{z}_i^{(\alpha)\mathrm{T}},\ldots,I_{\{k_i=K-1\}}\mathbf{z}_i^{(\alpha)\mathrm{T}}\right)^{\mathrm{T}}\right],
\end{aligned}$$

which, upon recognition of the form of $\hat{p}_{k_i}(\mathbf{z}_i^{(\alpha)}),$ gives for $k_i \le K-1$

$$\nabla_{\hat{\boldsymbol{\alpha}}}\hat{\omega}_i = s_i\pi_{k_i}\left[\left(\mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}, \mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_2^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}, \ldots, \mathbf{z}_i^{(\alpha)\mathrm{T}}\frac{\exp\left(\hat{\boldsymbol{\alpha}}_{K-1}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\boldsymbol{\alpha}}_{k_i}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)}\right)^{\mathrm{T}} - \frac{1}{\hat{p}_{k_i}(\mathbf{z}_i^{(\alpha)})}\left(I_{\{k_i=1\}}\mathbf{z}_i^{(\alpha)\mathrm{T}}, I_{\{k_i=2\}}\mathbf{z}_i^{(\alpha)\mathrm{T}}, \ldots, I_{\{k_i=K-1\}}\mathbf{z}_i^{(\alpha)\mathrm{T}}\right)^{\mathrm{T}}\right]. \tag{1}$$

For subjects with $k_i = K,$ we have

$$\begin{aligned}\nabla_{\hat{\boldsymbol{\alpha}}}\hat{\omega}_i &= s_i\pi_{k_i}\nabla_{\hat{\boldsymbol{\alpha}}}\left\{1+\sum\nolimits_{j=1}^{K-1}\exp\left(\hat{\boldsymbol{\alpha}}_j^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)\right\} \\ &= s_i\pi_{k_i}\left(\mathbf{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right), \mathbf{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\boldsymbol{\alpha}}_2^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right), \ldots, \mathbf{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\boldsymbol{\alpha}}_{K-1}^{\mathrm{T}}\mathbf{z}_i^{(\alpha)}\right)\right)^{\mathrm{T}}.\end{aligned} \tag{2}$$

Pugh *et al.* (1993) developed an estimator for the covariance matrix of the Cox model regression parameters when a weighted analysis is used to adjust the regression parameter estimates for missing covariates in some subjects. To make the adjustment, a logistic regression analysis is used to estimate the probability of missingness and a weight equal to the inverse of this probability for each subject is used in the Cox regression. The estimated covariance matrix of the Cox model parameter accounts for the variability in the logistic regression-based weight estimates. This situation, in which the weights derive from the probability of missingness, is completely analogous to the use of IPTW weights that derive from the probability of treatment assignment. Applying the results of Pugh *et al.* (1993) as described in Therneau and Grambsch (2000), the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated consistently, accounting for the variance in both the Cox regression and the logistic regression for the propensity score-based weights, by

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} = \mathbf{D}_{\hat{\boldsymbol{\beta}}}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{D}_{\hat{\boldsymbol{\alpha}}}\left(\mathbf{D}_{\hat{\boldsymbol{\alpha}}}^{\mathrm{T}}\mathbf{D}_{\hat{\boldsymbol{\alpha}}}\right)^{-1}\mathbf{D}_{\hat{\boldsymbol{\alpha}}}^{\mathrm{T}}\right)\mathbf{D}_{\hat{\boldsymbol{\beta}}},$$

where $\mathbf{D}_{\hat{\boldsymbol{\beta}}}$ is the matrix of dfbetas for $\hat{\boldsymbol{\beta}}$ and $\mathbf{D}_{\hat{\boldsymbol{\alpha}}}$ is the matrix of dfbetas for $\hat{\boldsymbol{\alpha}}$ from the logistic regression model used to estimate the probability of treatment assignment. The ith row of the dfbeta matrix closely approximates the change in the regression parameter estimate vector that would result from deleting subject i from the analysis set. Note that $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ is the covariance matrix of the residuals of the linear regression of the $\hat{\boldsymbol{\beta}}$ dfbeta on the $\hat{\boldsymbol{\alpha}}$ dfbeta, so the elements of $\hat{\boldsymbol{\beta}}$ estimated from the Cox regression model using the weights are asymptotically uncorrelated with the elements of $\hat{\boldsymbol{\alpha}}$.

The dfbetas for the Cox model can be computed from the score residuals and the Fisher information matrix. The score residual for subject i is

$$\mathbf{u}_i = \int_0^\infty \left\{ \mathbf{z}_i^{(\beta)}(t) - \overline{\mathbf{z}^{(\beta)}}(t) \right\} \mathrm{d}\, \hat{M}_i(t),$$

where

$$\overline{\mathbf{z}^{(\beta)}}(t) = \frac{\sum_{i=1}^n \hat{\omega}_i Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t)\right) \mathbf{z}_i^{(\beta)}(t)}{\sum_{i=1}^n \hat{\omega}_i Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t)\right)}$$

and

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(s)\right) \mathrm{d}\, \hat{\Lambda}_0(s)$$

is the martingale residual process (Therneau and Grambsch 2000). Here $N_i(t)$ is the event-counting process for subject i, $Y_i(t)$ is the indicator that subject i is in the risk set (still being followed) at time t, and

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n \hat{\omega}_i \, \mathrm{d}\, N_i(s)}{\sum_{i=1}^n \hat{\omega}_i Y_i(s) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(s)\right)}$$

is the baseline cumulative hazard estimator. The ith row of the dfbeta matrix $\mathbf{D}_{\hat{\boldsymbol{\beta}}}$ is $\hat{\omega}_i \mathbf{u}_i^{\mathrm{T}} \hat{\mathbf{I}}^{-1}(\hat{\boldsymbol{\beta}})$, where $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})$ is the estimated Fisher information matrix evaluated at the maximum partial likelihood estimate. $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})$ is the matrix of the negatives of the second derivatives of the log partial likelihood with respect to the regression parameter estimates and its inverse is output by standard proportional hazard regression packages as the (naïve) model-based estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

Czepiel (2002) showed that for the multinomial logistic regression model, the ith row of the dfbeta matrix $\mathbf{D}_{\boldsymbol{\alpha}}$ is given by

$$\left(\left\{ I_{\{k_i=1\}} - \hat{p}_1(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}}, \left\{ I_{\{k_i=2\}} - \hat{p}_2(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}}, \ldots, \left\{ I_{\{k_i=K-1\}} - \hat{p}_{K-1}(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}} \right) \hat{\mathbf{V}}_{\hat{\boldsymbol{\alpha}}},$$

where $I_{\{k_i=k\}}$ is the indicator for whether subject i received treatment k and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\alpha}}}$ is the estimated covariance matrix of the maximum likelihood logistic regression parameter estimator $\hat{\boldsymbol{\alpha}}$, that is,

the inverse of the Fisher information matrix. With stratified cohort sampling weighting, in which subject i is weighted by s_i, the inverse of the sampling fraction for that subject's stratum, the ith row becomes

$$s_i \left(\left\{ I_{\{k_i=1\}} - \hat{p}_1(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}}, \left\{ I_{\{k_i=2\}} - \hat{p}_2(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}}, \ldots, \left\{ I_{\{k_i=K-1\}} - \hat{p}_{K-1}(\mathbf{z}_i^{(\alpha)}) \right\} \mathbf{z}_i^{(\alpha)\mathrm{T}} \right) \hat{\mathbf{V}}_{\hat{\boldsymbol{\alpha}}},$$

where $\hat{\mathbf{V}}_{\hat{\boldsymbol{\alpha}}}$ is now the (naïve) model-based estimate of the covariance matrix of $\hat{\boldsymbol{\alpha}}$ in the logistic regression using the cohort sampling weights.

The estimated cumulative hazard at time *t* for a patient with covariate vector $\mathbf{z}^{(\beta)}$ is

$$\hat{\Lambda}(t; \mathbf{z}^{(\beta)}) = \int_0^t \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}^{(\beta)}(s) \right) \mathrm{d}\, \hat{\Lambda}_0(s).$$

The gradient of the increment in the baseline hazard at time *t* estimator with respect to the Cox regression parameter estimate vector is

$$\nabla_{\hat{\boldsymbol{\beta}}}\, \mathrm{d}\, \hat{\Lambda}_0(t) = -\overline{\mathbf{z}^{(\beta)}}(t)\, \mathrm{d}\, \hat{\Lambda}_0(t).$$

Thus,

$$\begin{aligned} \nabla_{\hat{\boldsymbol{\beta}}} \hat{\Lambda}(t; \mathbf{z}^{(\beta)}) &= \int_0^t \left\{ \nabla_{\hat{\boldsymbol{\beta}}} \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}^{(\beta)}(s) \right) \right\} \mathrm{d}\, \hat{\Lambda}_0(s) + \int_0^t \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}^{(\beta)}(s) \right) \nabla_{\hat{\boldsymbol{\beta}}}\, \mathrm{d}\, \hat{\Lambda}_0(s) \\ &= \int_0^t \left\{ \mathbf{z}^{(\beta)}(s) - \overline{\mathbf{z}^{(\beta)}}(s) \right\} \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}^{(\beta)}(s) \right) \mathrm{d}\, \hat{\Lambda}_0(s) \end{aligned}$$

The gradient of the increment in baseline hazard estimator at time *t* with respect to the logistic regression parameter estimate vector $\hat{\boldsymbol{\alpha}}$ is

$$\begin{aligned} \nabla_{\hat{\boldsymbol{\alpha}}}\, \mathrm{d}\, \hat{\Lambda}_0(t) &= \nabla_{\hat{\boldsymbol{\alpha}}} \left\{ \frac{\sum_{i=1}^n \hat{\omega}_i\, \mathrm{d}\, N_i(t)}{\sum_{i=1}^n \hat{\omega}_i Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t) \right)} \right\} \\ &= \frac{\sum_{i=1}^n \left(\nabla_{\hat{\boldsymbol{\alpha}}} \hat{\omega}_i \right) \mathrm{d}\, N_i(t)}{\sum_{i=1}^n \hat{\omega}_i Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t) \right)} \\ &\quad - \frac{\left\{ \sum_{i=1}^n \hat{\omega}_i\, \mathrm{d}\, N_i(t) \right\} \sum_{i=1}^n \left(\nabla_{\hat{\boldsymbol{\alpha}}} \hat{\omega}_i \right) Y_i(s) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t) \right)}{\left\{ \sum_{i=1}^n \hat{\omega}_i Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i^{(\beta)}(t) \right) \right\}^2}, \end{aligned}$$

where $\nabla_{\hat{\boldsymbol{\alpha}}} \hat{\omega}_i$ is given in equations (1) and (2). This lets us compute

$$\nabla_{\hat{\boldsymbol{\alpha}}} \hat{\Lambda}(t; \mathbf{z}^{(\beta)}) = \int_0^t \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}^{(\beta)}(s) \right) \nabla_{\hat{\boldsymbol{\alpha}}}\, \mathrm{d}\, \hat{\Lambda}_0(s).$$

We also need to account for the variability in the number and timing of jumps in the overall event counting process. Let T_i be the end of follow-up for subject i, at which time either an event occurred $\left(\mathrm{d}\,N_i(T_i)=1\right)$ or the subject's time to event was censored $\left(\mathrm{d}\,N_i(T_i)=0\right)$. Then we have

$$\frac{\partial\hat{\Lambda}\left(t;\mathbf{z}^{(\beta)}\right)}{\partial\,\mathrm{d}\,N_i(T_i)}=I_{\{T_i\le t\}}\exp\left\{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}^{(\beta)}(T_i)\right\}\mathrm{d}\,\hat{\Lambda}_0(T_i).$$

The variance of $\hat{\Lambda}(t;\mathbf{z})$ can then be consistently estimated by

$$\widehat{\mathrm{Var}}\left\{\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}=\left\{\nabla_{\hat{\boldsymbol{\beta}}}\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}^{\mathrm{T}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}\left\{\nabla_{\hat{\boldsymbol{\beta}}}\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}+\left\{\nabla_{\hat{\boldsymbol{\alpha}}}\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}^{\mathrm{T}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\alpha}}}\left\{\nabla_{\hat{\boldsymbol{\alpha}}}\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}$$
$$+\sum_{i=1}^{n}\left\{\frac{\partial\hat{\Lambda}\left(t;\mathbf{z}^{(\beta)}\right)}{\partial\,\mathrm{d}\,N_i(T_i)}\right\}^2\mathrm{d}\,N_i(T_i).$$

The variance of the log cumulative hazard estimator $\hat{\rho}(t;\mathbf{z})=\ln\hat{\Lambda}(t;\mathbf{z})$ is estimated consistently by

$$\widehat{\mathrm{Var}}\left\{\hat{\rho}(t;\mathbf{z}^{(\beta)})\right\}=\frac{\widehat{\mathrm{Var}}\left\{\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}}{\left\{\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right\}^2}.$$

Since the risk at time *t* is estimated as

$$\hat{r}(t;\mathbf{z}^{(\beta)})=1-\exp\left(-\hat{\Lambda}(t;\mathbf{z}^{(\beta)})\right)=1-\exp\left[-\exp\left\{\hat{\rho}(t;\mathbf{z}^{(\beta)})\right\}\right],$$

a level α confidence interval for the risk at time *t* has endpoints

$$1-\exp\left[-\exp\left\{\hat{\rho}(t;\mathbf{z}^{(\beta)})\pm\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\widehat{\mathrm{Var}}\left(\hat{\rho}(t;\mathbf{z}^{(\beta)})\right)\right\}\right],$$

where Φ is the cumulative distribution function of the standard normal distribution.

In some situations, it may be appropriate to fit a Cox proportional hazards regression allowing a distinct baseline hazard function in each stratum of the population. If S_k is the subset of patients in stratum k, the baseline cumulative hazard function estimate for stratum k is

$$\hat{\Lambda}_0(t)=\int_0^t\frac{\sum_{i=1}^{n}I_{\{i\in S_k\}}\hat{\omega}_i\,\mathrm{d}\,N_i(s)}{\sum_{i=1}^{n}I_{\{i\in S_k\}}\hat{\omega}_iY_i(s)\exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i^{(\beta)}(s)\right)}.$$

Macro Risk_Est_PH_reg_prop_score

Macro Risk_Est_PH_reg_prop_score computes risk estimates from a proportional hazards regression with IPTW weighting using the methods described above. The macro is called as follows:

```
%Risk_Est_PH_reg_time_dep(
    /* Input Specification */  indsn=, byvar= , vars=, vars_logistic=,
                    time=, censor=, censorlist=,
                    entrytime=, weight=,
                    response=, stabilize=, truncate_pct=,
                    programming_statements=%str(), calc_vars=, covariate_dsn=,
    /* Analysis Parameters */  risk_time=, print_phreg=, print_logistic=,
                                alpha=, strata=, CI_method=
    /* Output Specification */ outdsn=, Risk=, Risk_LCL=, Risk_UCL=,
                        CumHaz, CumHaz_LCL=, CumHaz_UCL=,
                        LogCumHaz=, SE_LogCumHaz=,
                        Parameter_estout=, parameter_covout=,
                    std_diff_vars=,
                    std_diff_out=, std_diff_graph_name= ,graph_path=
        );
```

The macro parameters are defined in Table 1. The time dependence of the covariates, if any, is defined using programming statements (as are used for PROC PHREG). It is assumed that these programming statements, when applied to the input data set (and the covariate data set, if specified), will uniquely determine the covariate values at the time specified by the input data set variable given by the macro parameter time. Note the programming statements will need to be enclosed in %str() so that the semicolon(s) will not cause a syntax error.

Table 1. Macro Risk_Est_PH_reg_prop_score Parameters

Parameter	Type	Required?	Default Value	Description
indsn	$	Yes	(at temporary library)	(Libname reference and) file name containing input data set.
byvar	$	No	—	Optional list of variables to do the analysis by.

Table 1. Macro Risk_Est_PH_reg_prop_score Parameters

Parameter	Type	Required?	Default Value	Description
vars	$	Yes	—	List of input data set variables to be used as the covariate in the Cox model used to estimate the risk. If the programming statements create the variables that are to be included in the model, list the variables thus created along with any time-invariant covariates.
vars_logistic	#	Yes	—	List of input data set variables to be used as the covariates in the logistic regression model for estimating the propensities.
time	#	Yes	—	Input data set variable containing the time to event (or censoring).
censor	#	Yes	—	Input data set variable indicating whether the observed time to event was censored.
censorlist	#	No	0	List of values of variable censor that indicate a censored observation. Default is the single value 0.
entrytime	#	No	—	Optional input data set variable containing a left truncation time for each observation.
weight	#	No	—	Input data set variable giving the observation's sampling weight if cohort sampling was used.
response	$/#	Yes	—	Input data set variables giving the response for which the propensity will be estimated (such as treatment or biomarker use). This may be a binary or multinomial outcome.
stabilize	$	No	yes	If this parameter is set to yet, stabilized propensity score weights will be used (multiplying the inverse probability by the proportion of patients with the response).
truncate_pct	#	No	0	If this parameter is set to a non-negative number, the propensity score weights will be truncated the &truncate_pct and 100-&truncate_pct percentiles. The percentage will be rounded to the nearest tenth of a percent. The percentage specified must be non-negative and less than 50. A typical value is 5 (truncating at the 5th and 95th percentiles).
programming_ statements	$	No	—	%str()-enclosed text string including programming statements that will be inserted into proc PHREG and various data steps to compute the time-dependent covariate values. For example: programming_statements = %str(if time <= 3 then x_3 = 0; else x_3 = x;) If no programming statements are entered, the risk calculations will be made for covariates that are constant over time.

Table 1. Macro Risk_Est_PH_reg_prop_score Parameters

Parameter	Type	Required?	Default Value	Description
calc_vars	$	Yes, if time-dependent covariates are used	—	List of variables that are used in the calculation of the time-dependent covariate values. Include all non-time-dependent covariate values in this list, too. Leave this parameter blank if time dependent variables are not used.
covariate_dsn	$	No	(input data set)	(Libname reference and) the name of a data set that contains the covariate values for which the risk is to be estimated. The data set must have all the variables included in the model, or that are required to derive these variables if the model has time-dependent covariates derived using programming statements. The data set must also include the stratification variable if the model is stratified. If no covariate data set is specified, the risk will be estimated for every patient in the main input data set.
risk_time	#	Yes	—	This is the time at which the risk is assessed for each patient. That is, the risk is defined as the probability that the patient will have the event on or before risk_time.
print_phreg	$	No	yes	If this parameter is set to no, output from the PHREG model fit will not be printed. The variability estimates and confidence intervals will account for the weight estimation variability.
print_logistic	$	No	yes	If this parameter is set to no, the PROC LOGISTIC output will not be printed.
alpha	#	No	0.05	The macro will compute a 100(1-alpha)% confidence interval for the risk and cumulative hazard.
strata	$/#	No	—	Character string giving input data set variable by which the proportional hazards regression analysis will be stratified.
CI_method	$	No	loglog	Character string giving the method for computing the confidence intervals. If linear is specified, the confidence interval is computed on the risk scale. If log is specified, the confidence interval is computed on the cumulative hazard scale and transformed to the risk scale. If loglog is specified, the confidence interval is computed on the log cumulative hazard scale and transformed to the risk scale.
outdsn	$	Yes	(at temporary library)	(Libname reference and) the output data set name. This data set will contain all the records and variables of the covariate data set (or the input data set if no separate coavariate data set is specified) plus the variables named by the following eight macro parameters.

Table 1. Macro Risk_Est_PH_reg_prop_score Parameters

Parameter	Type	Required?	Default Value	Description
Risk	#	No	Risk	Name of output data set variable that will contain the risk estimate.
Risk_LCL	#	No	Risk_LCL	Name of output data set variable will contain the lower limit of a 1-alpha confidence interval for the risk.
Risk_UCL	#	No	Risk_UCL	Name of output data set variable will contain the upper limit of a 1-alpha confidence interval for the risk.
CumHaz	#	No	CumHaz	Name of output data set variable that will contain the cumulative hazard estimate.
CumHaz_LCL	#	No	CumHaz_LCL	Name of output data set variable will contain the lower limit of a 1-alpha confidence interval for the cumulative hazard.
CumHaz_UCL	#	No	CumHaz_UCL	Name of output data set variable will contain the upper limit of a 1-alpha confidence interval for the cumulative hazard.
LogCumHaz	#	No	LogCumHaz	Name of output data set variable that will contain the log cumulative hazard estimate.
SE_LogCumHaz	#	No	SE_LogCumHaz	Name of output data set variable that will contain the estimated standard error of the log cumulative hazard estimate.
Parameter_estout	$	No	(at temporary library)	Optional libname reference and data set name that will contain the Cox regression parameter estimates, standard errors, chi-square statitistics, p-values, hazard ratio estimates and confidence intervals, all computed using the covariance matrices accounting for variability in the propensity score-based weight estimates.
Parameter_covout	$	No	(at temporary library)	Optional libname reference and name of data set that will contain the Cox regression parameter estimate covariance matrix.
std_diff_vars	$	No	—	Optional list of variables for which the standardized differences among logistic| regression outcomes will be assessed.
Std_diff_out	$	No	—	Optional libname reference and name of data set that will contain the weighted and| unweighted standardized differences among the logistic regression response outcomes for each logistic regression covariate.
Std_diff_graph_name	$	No	—	Optional name for standardized difference graphs. The graphs will be placed in| the directory named in parameter graph_path and suffixes of the form i_j will be affixed to the graph name specifying the number logistic regression outcomes being compared in each graph.
Graph_path	$	No	—	Optional specification of folder in which to place the standardized difference graphs.

References

Austin PC, Stuart EA (2015). Moving towards best practiced when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* **34**:3661–3679.

Cole SR, Hernán MA (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168**:656–664.

Czepiel SA (2002). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.* http://czep.net

Pugh M, Robbins J, Lipsitz S, Harrington D (1993). *Inference in the Cox proportional hazards model with missing covariates.* Technical Report 758Z. Department of Biostatistics, Harvard School of Public Health.

Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **76**:41–55.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.