Documentation
`sas-twophase-package`:
Extended pseudo likelihood estimation with the `extPL`-Macro

Author: **D. Enders**
Date: **06.03.2017**
Version: **1**

Contents

1 General

The purpose of the SAS-macro extPL is the conduct of logistic regression analyses with data in a two-phase setting. See the documentation TPDocu.pdf and TPMethods.pdf of the %twophase-Macro for details on the standard methods. This macro implements the pseudo likelihood method of Scott and Wild (2011) by assuming a logistic regression model for participation in phase 2. The method naturally extends the pseudo likelihood method of Breslow and Cain (1988).

2 Usage

The usage of the extPLmacro is as follows:

```
%extPL{   path_in    = ,
          sel_var    = ,
          out_var    = ,
          caco       = ,
          ph2_ind    = ,
          path_out   = ,
          diag       = }
```

3 Arguments

The following parameters can be set:

Parameter	Description
path_in	**Path of the input dataset. The value is of the form path_in=*libname.filename*. If the dataset is located in the workspace, libname can be omitted. Value: string Default: none**
sel_var	**Names of the regression variables in the logistic regression model for participation in phase 2. The names must be separated by blanks. Value: string Default: none**
out_var	**Names of the regression variables in the logistic regression model for the outcome. The names must be separated by blanks. Value: string Default: none**

`caco`	**Name of outcome (case-control) variable. The variable must have two values: 1 (cases); 0 (controls)** **Value: string** **Default: None**
`path_out`	**Path of the output dataset. The value is of the form** `path_out=`***libname.filename*****. If the dataset should be located in the workspace, libname can be omitted.** **Value: string** **Default: none**
`ph2_ind`	**Name of variable indicating participation in phase 2. The variable must have two values: 1 (participating in phase 2); 0 (not participating in phase 2)** **Value: string** **Default: None**
`diag`	**If not set to zero, details of SAS procedures on the logistic regression analyses and details on the extrapolation are directed to the SAS-output.** **Value: numeric** **Default: 0**

4 Input data

The input data contains all phase 1 and phase 2 variables of the study. Patients participating in phase 2 are identified by `&ph2_ind=1`**. Phase 2 variables for patients with** `&ph2_ind=0` **are ignored an can thus have missing values.**
The macro uses the variables `&caco`**,** `&ph2_ind`**,** `&sel_var` **and** `&out_var`**. The variables specified in** `&sel_var` **and** `&out_var` **enter in exactly this form into the linear predictor of the logistic regression model for the participation in phase 2 and for the outcome, respectively.** `&sel_var` **can only include phase 1 variables, while** `&out_var` **might include factors of phase 1 and phase 2.**
`extPL` **has no capabilities to treat categorical variables as factors or to build interaction variables. This means that all data manipulation tasks like construction of dummies or building interaction terms have to be executed before the macro is called.**

5 Estimation process

The process of estimation is as follows:

- **The phase 2 data (**`&ph2_ind=1`**) is extracted from the input data in** `&path_in`**. The number of observations in phase 2 are stored in** N_{Ph2}**.**

- **A logistic regression model for participation in phase 2, separately for cases and controls, is fitted to the phase 1 data using the model**

  $$\text{logit}(\Pr(\texttt{\&ph2\_ind} = 1|\texttt{\&sel\_var}, \texttt{\&caco} = 1) = (1, \texttt{\&sel\_var})\hat{\alpha}_1 \text{ and}$$
  $$\text{logit}(\Pr(\texttt{\&ph2\_ind} = 1|\texttt{\&sel\_var}, \texttt{\&caco} = 0) = (1, \texttt{\&sel\_var})\hat{\alpha}_0.$$

  The parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_0$ and the respective estimated covariance matrices $\widehat{Cov}(\hat{\alpha}_1)$ and $\widehat{Cov}(\hat{\alpha}_0)$ are stored.

- **For each patient in phase 2, the offset**

  $$\omega = \frac{\widehat{\Pr}(\texttt{\&ph2\_ind} = 1|\texttt{\&sel\_var}, \texttt{\&caco} = 1)}{\widehat{\Pr}(\texttt{\&ph2\_ind} = 1|\texttt{\&sel\_var}, \texttt{\&caco} = 0)} = \frac{\frac{\exp((1,\texttt{\&sel\_var})\hat{\alpha}_1)}{1+\exp((1,\texttt{\&sel\_var})\hat{\alpha}_1)}}{\frac{\exp((1,\texttt{\&sel\_var})\hat{\alpha}_0)}{1+\exp((1,\texttt{\&sel\_var})\hat{\alpha}_0)}} \quad (5.1)$$

  is calculated. Note that this includes the estimation of the probability for participation in phase 2, if the patient would have been a case and if the patient would have been a control. The calculation of ω thus involve extrapolation: If the patient is a control and its covariate pattern is only present in controls and not in cases, the calculation of $\widehat{\Pr}(\texttt{\&ph2\_ind} = 1|\texttt{\&sel\_var}, \texttt{\&caco} = 1)$ is an extrapolation of the observed data for cases to the covariate pattern of this patient. Obviously, this extrapolation occurs also in patients, which are cases.

- **A logistic regression for the outcome `&caco` is conducted, using the following pseudo model with ω as offset:**

  $$\text{logit}(\Pr^*(\texttt{\&caco} = 1|\texttt{\&out\_var}, \texttt{\&ph2\_ind} = 1) = (1, \texttt{\&out\_var})\beta + \omega. \quad (5.2)$$

  The parameter estimates $\hat{\beta}$ and the respective covariance matrix $\widehat{Cov}(\hat{\beta})$ are stored.
  Note that $\widehat{Cov}(\hat{\beta})$ do not account for the estimation of α_0 and α_1, which led to a robust covariance estimator (Scott and Wild, 2011). A consistent covariance estimator, presented in a general form in Scott and Wild (2011)[Equation (5)], is calculated by

  $$\begin{aligned}\widehat{Cov}^*\left(\hat{\beta}\right) =& \widehat{Cov}(\hat{\beta}) \\ &- \widehat{Cov}(\hat{\beta}) I_{01,1} \widehat{Cov}\left(\hat{\alpha}_1\right) I_{01,1}^t \widehat{Cov}(\hat{\beta}) \\ &- \widehat{Cov}(\hat{\beta}) I_{01,0} \widehat{Cov}\left(\hat{\alpha}_0\right) I_{01,0}^t \widehat{Cov}(\hat{\beta}),\end{aligned} \quad (5.3)$$

  with $I_{01,1} = X_\beta^t Diag(\mu\#(1-\mu))S_{\alpha_1}$ and $I_{01,0} = X_\beta^t Diag(\mu\#(1-\mu))S_{\alpha_0}$ (# denotes rowwise scalar multiplication). Here, X_β is the design matrix of the variables `&out_var` for patients in phase 2, μ the vector of

patient specific outcome probabilities under the pseudo model (5.2)**, i.e.**

$$\boldsymbol{\mu} = \exp\left(\boldsymbol{X}_{\beta}\hat{\boldsymbol{\beta}} + \boldsymbol{\omega}\right) \# \left(1 + \exp\left(\boldsymbol{X}_{\beta}\hat{\boldsymbol{\beta}} + \boldsymbol{\omega}\right)\right)^{-1}$$

and $\boldsymbol{S}_{\alpha_i} = \boldsymbol{X}_{\alpha}\#(1-\boldsymbol{\mu}_{\alpha_i})$**, where** $\boldsymbol{X}_{\alpha}$ **is the design matrix of the variables** `&sel_var` **for patients in phase 2 and**

$$\boldsymbol{\mu}_{\alpha_i} = \frac{\exp\left(\boldsymbol{X}_{\alpha}\hat{\boldsymbol{\alpha}}_i\right)}{1 + \exp\left(\boldsymbol{X}_{\alpha}\hat{\boldsymbol{\alpha}}_i\right)},\ i = 0, 1.$$

6 Output data

The output data set is stored under `path_out` **and contains the following variables:** `Variable` **(One row for the intercept and each element of** `&out_var`**),** `Est` **(the elements of** $\hat{\beta}$**),** `Stderr` **(square root of the diagonal elements of** $\widehat{Cov}^{*}(\hat{\beta})$**) and** `Stderr_rob` **(square root of the diagonal elements of** $\widehat{Cov}(\hat{\beta})$**). Note that the estimate of the intercept is only interpretable in prospective logistic regression (Schill and Drescher, 1997).**
Details of both logistic regression analyses in Section 5 as well as the number of patterns with extrapolations in the calculation of ω **can be directed to the SAS output by specifying** `&diag=1`**.**

References

N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75:11–20, 1988.

W. Schill and K. Drescher. Logistic analysis of studies with two-stage sampling: A comparison of four approaches. *Statistics in Medicine*, 16:117–132, 1997.

A. J. Scott and C. J. Wild. Fitting regression models with response-biased samples. *The Canadian Journal of Statistics*, 39:519–536, 2011.