

CoxRegression

Module name: CoxRegression
Description: Cox regression using clinical data and .cls file
Author: Yujin Hoshida (Broad Institute) gp-help@broad.mit.edu
Date: 5/20/2008
Release: 1.0

Summary: Once a genomics marker or predictive model for clinical outcome is built, the next step is to test whether it really associates with outcome of new cohorts, or whether it has better predictive ability compared to already known clinical markers/predictive models. "Survival analysis" modules perform such analyses seamlessly after clustering/marker gene/prediction analyses.

Cox proportional hazard modeling (Cox regression) is commonly used to assess association of variable(s) of interest with time-to-event data (e.g., death, tumor recurrence,...) [1-3]. For the CoxRegression module, a .cls file identifies the phenotype label for each sample and a separate tab-delimited text file defines all other clinical variables. The module generates a result file similar to the one shown below. Interpreting the results requires basic knowledge of biostatistics and survival data analysis.

```
# Cox regression results
tie data handling: Efron approximation
variable selection: none
Call:
coxph(formula = Surv(time, censor) ~ factor(cls), data = surv.data,
      method = "efron", robust = F)

n= 100

      coef exp(coef) se(coef)  z      p
factor(cls)2  1.93      6.86  0.364  5.3 1.2e-07
      exp(coef) exp(-coef) lower .95 upper .95
factor(cls)2    6.86      0.146   3.36    14

Rsquare= 0.331 (max possible= 0.988 )
Likelihood ratio test= 40.2 on 1 df, p=2.3e-10
Wald test          = 28.1 on 1 df, p=1.18e-07
```

Score (logrank) test = 37.5 on 1 df, p=8.9e-10

Reference:

1. Therneau TM. *Modeling Survival Data: Extending the Cox Model*. 2001, Springer, Berlin
2. Lumley T. *survival 2.20*, R package, <http://www.r-project.org/>
3. Venables WN, Ripley BD. *Survival Analysis in Modern Applied Statistics with S-PLUS*, 3rd Ed. 1999, Springer, New York.

Parameters:

| Name | Description | Choices |
|----------------------------|--|-----------------|
| input surv data filename | clinical data containing survival time and censor status - .txt Format shown below. | |
| input cls file | class information file -.cls Format described in GenePattern File Formats . | |
| output file | Name for the output file. | |
| time | Column name for the survival time field. Must be a column in the input surv data file. | Default: time |
| status | Column name for the censor field (event="1", no event="0"). Must be a column in the input surv data file. | Default: status |
| variable continuous | Comma separated list of the continuous dependent variables (e.g. age) to use for the analysis. Each variable must be a column in the input surv data file. | |
| variable category | Comma separated list of the categorical dependent variables (e.g. gender) to use for the analysis. Each variable must be a column in the input surv data file or cls (to use the classes in the input cls file). | |
| variable interaction terms | Separate interaction variables with a | |

| | | |
|--------------------|--|--|
| | colon (e.g. sex:cls). Separate sets of interaction variables with a comma (e.g. sex:cls, sex:grade). Each variable must be a column in the input surv data file or cls (to use the classes in the input cls file). | |
| strata | Fit is done for each strata. Use the default value unless you are familiar with clinical data analysis. | |
| input subgroup | Analyze a subgroup of the data. Identify the subgroup by specifying a variable followed by a colon and a value for the variable. For example: gender:M or cls:1). | |
| variable selection | Select the way dependent variables are entered into the model. Use the default value unless you are familiar with clinical data analysis. | Default: no Stepwise selection Forward selection Backward selection |

Format of input clinical data file:

Tab-delimited text file. The first row contains column heads. The first column must be the sample identifier; the column head is your choice; the order of samples in this file and the input cls file must match. Each subsequent column contains a clinical variable. For example:

| name | time | sensor | cls.clinical | labdata1 | labdata1 | age | sex | grade |
|-------|------|--------|--------------|----------|----------|-----|-----|-------|
| case1 | 207 | 0 | poor | 0.9 | 430 | 72 | 1 | 4 |
| case4 | 49 | 0 | poor | 2.7 | 220 | 54 | 1 | 4 |
| case5 | 267 | 0 | poor | 28.7 | 500 | 53 | 1 | 4 |
| case7 | 8 | 0 | poor | 2.4 | 1090 | 75 | 0 | 4 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Identify continuous (e.g., age) or categorical (e.g., sex) variables for the analysis by entering the column heads in the *variable continuous* or *variable category* fields,

respectively. Identify the survival time and censor columns for the analysis by entering the column heads “time” and “censor” in the *time* and *status* fields, respectively.

Example input files: [surv.txt](#), [surv.cls](#)

The example files contain mock data. To run an analysis using these files, enter the column head “time” in the *time* field, “censor” in the *status* field, and “cls” in the *variable category* field.

Output files:

CoxRegression_result.txt: Summary of the results.

Platform dependencies:

task type: Survival Analysis
CPU type: any
OS: any
Language: R (version 2.0.1 or above)
Support files: none