Package 'cvGEE'

July 23, 2019

Title Cross-Validated Predictions from GEE

Version 0.3-0

Date 2019-07-20

Maintainer Dimitris Rizopoulos <d.rizopoulos@erasmusmc.nl>

BugReports https://github.com/drizopoulos/cvGEE/issues

Description Calculates predictions from generalized estimating equations and internally cross-validates them using the logarithmic, quadratic and spherical proper scoring rules; Kung-Yee Liang and Scott L. Zeger (1986) <doi:10.1093/biomet/73.1.13>.

Suggests geepack, lattice, knitr, rmarkdown, pkgdown

Encoding UTF-8

LazyLoad yes

LazyData yes

License GPL (>= 3)

URL https://drizopoulos.github.io/cvGEE/,
https://github.com/drizopoulos/cvGEE

VignetteBuilder knitr

RoxygenNote 6.1.1

NeedsCompilation no

Author Dimitris Rizopoulos [aut, cre]
(<https://orcid.org/0000-0001-9397-0900>)

Repository CRAN

Date/Publication 2019-07-23 14:52:05 UTC

R topics documented:

aids *Didanosine versus Zalcitabine in HIV Patients*

Description

A randomized clinical trial in which both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs in treating patients who had failed or were intolerant of zidovudine (AZT) therapy.

Format

A data frame with 1408 observations on the following 9 variables.

`patient` patients identifier; in total there are 467 patients.

`Time` the time to death or censoring.

`death` a numeric vector with 0 denoting censoring and 1 death.

`CD4` the CD4 cells count.

`obstime` the time points at which the CD4 cells count was recorded.

`drug` a factor with levels `ddC` denoting zalcitabine and `ddI` denoting didanosine.

`gender` a factor with levels `female` and `male`.

`prevOI` a factor with levels `AIDS` denoting previous opportunistic infection (AIDS diagnosis) at study entry, and `noAIDS` denoting no previous infection.

`AZT` a factor with levels `intolerance` and `failure` denoting AZT intolerance and AZT failure, respectively.

Note

The data frame `aids.id` contains the first CD4 cell count measurement for each patient. This data frame is used to fit the survival model.

References

Goldman, A., Carlin, B., Crane, L., Launer, C., Korvick, J., Deyton, L. and Abrams, D. (1996) Response of CD4+ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **11**, 161–169.

Guo, X. and Carlin, B. (2004) Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician* **58**, 16–24.

cvGEE *Proper Scoring Rules for Generalized Estimating Equations*

Description

Calculates the logarithmic, quadratic/Brier and spherical scoring rules based on generalized estimation equations.

Details

Package:	cvGEE
Type:	Package
Version:	0.3-0
Date:	2019-07-20
License:	GPL (>=3)

The package provides the estimated values of the scoring rules for each observation of the original dataset. These values can be summarized/averaged or used in figures to evaluate how the GEE performs in different ranges of the data.

Author(s)

Dimitris Rizopoulos

Maintainer: Dimitris Rizopoulos <d.rizopoulos@erasmusmc.nl>

References

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis* **13**, 223-242. doi:10.1287/deca.2016.0337

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22. doi:10.1093/biomet/73.1.13

cv_gee *Proper Scoring Rules for Generalized Estimating Equations*

Description

Calculates the logarithmic, quadratic/Brier and spherical scoring rules based on generalized estimation equations.

Usage

```
cv_gee(object, rule = c("all", "quadratic", "logarithmic", "spherical"),
  max_count = 500, K = 5L, M = 10L, seed = 1L, return_data = FALSE)
```

Arguments

object	an object inheriting from class "geeglm" of the **geepack**.
rule	character string indicating the type of scoring rule to be used.
max_count	numeric scalar or vector denoting the maximum count up to which to calculate probabilities; this is relevant for count response data.
K	numeric scalar indicating the number of folds used in the cross-validation procedure.
M	numeric scalar denoting how many times the split of the data in K folds will be performed. The reported scoring rules values are the average over the M replicates.
seed	numeric scalre providing the seed used in the procedure.
return_data	logical; if TRUE the values of the scoring rules are ruturned as extra columns of the data behind object.

Value

A list or a data.frame with elements or (extra) columns the values of the logarithmic, quadratic and spherical scoring rules calculated based on the GEE object.

Author(s)

Dimitris Rizopoulos <d.rizopoulos@erasmusmc.nl>

References

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis* **13**, 223-242. doi:10.1287/deca.2016.0337

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22. doi:10.1093/biomet/73.1.13

Examples

```
library("geepack")
library("lattice")

pbc2$serBilirD <- as.numeric(pbc2$serBilir > 1.2)

fm1 <- geeglm(serBilirD ~ year, family = binomial(), data = pbc2,
              id = id, corstr = "exchangeable")

fm2 <- geeglm(serBilirD ~ year * drug, family = binomial(), data = pbc2,
              id = id, corstr = "exchangeable")

plot_data <- cv_gee(fm1, return_data = TRUE, M = 5)
plot_data$model_year <- plot_data$.score
plot_data$model_year_drug <- unlist(cv_gee(fm2, M = 5))

xyplot(model_year + model_year_drug ~ year | .rule, data = plot_data,
```

```
        type = "smooth", auto.key = TRUE, layout = c(3, 1),
        scales = list(y = list(relation = "free")),
        xlab = "Follow-up time (years)", ylab = "Scoring Rules")
```

---

pbc2 *Mayo Clinic Primary Biliary Cirrhosis Data*

Description

Followup of 312 randomised patients with primary biliary cirrhosis, a rare autoimmune liver disease, at Mayo Clinic.

Format

A data frame with 1945 observations on the following 20 variables.

`id` patients identifier; in total there are 312 patients.

`years` number of years between registration and the earlier of death, transplantion, or study analysis time.

`status` a factor with levels `alive`, `transplanted` and `dead`.

`drug` a factor with levels `placebo` and `D-penicil`.

`age` at registration in years.

`sex` a factor with levels `male` and `female`.

`year` number of years between enrollment and this visit date, remaining values on the line of data refer to this visit.

`ascites` a factor with levels `No` and `Yes`.

`hepatomegaly` a factor with levels `No` and `Yes`.

`spiders` a factor with levels `No` and `Yes`.

`edema` a factor with levels `No edema` (i.e., no edema and no diuretic therapy for edema), `edema no diuretics` (i.e., edema present without diuretics, or edema resolved by diuretics), and `edema despite diuretics` (i.e., edema despite diuretic therapy).

`serBilir` serum bilirubin in mg/dl.

`serChol` serum cholesterol in mg/dl.

`albumin` albumin in gm/dl.

`alkaline` alkaline phosphatase in U/liter.

`SGOT` SGOT in U/ml.

`platelets` platelets per cubic ml / 1000.

`prothrombin` prothrombin time in seconds.

`histologic` histologic stage of disease.

`status2` a numeric vector with the value 1 denoting if the patient was dead, and 0 if the patient was alive or transplanted.

Note

The data frame `pbc2.id` contains the first measurement for each patient. This data frame is used to fit the survival model.

References

Fleming, T. and Harrington, D. (1991) *Counting Processes and Survival Analysis*. Wiley, New York.

Therneau, T. and Grambsch, P. (2000) *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.