Package 'pldamixture'

March 5, 2024

Type Package

Title Post-Linkage Data Analysis Based on Mixture Modelling

Version 0.1.0

Depends R (>= 3.5.0)

Imports stats, survival

Description

Perform inference in the secondary analysis setting with linked data potentially containing mismatch errors. Only the linked data file may be accessible and information about the record linkage process may be limited or unavailable. Implements the 'General Framework for Regression with Mismatched Data' developed by Slawski et al. (2023) <arXiv:2306.00909>. The framework uses a mixture model for pairs of linked records whose two components reflect distributions conditional on match status, i.e., correct match or mismatch. Inference is based on composite likelihood and the Expectation-Maximization (EM) algorithm. The package currently supports Cox Proportional Hazards Regression (right-censored data only) and Generalized Linear Regression Models (Gaussian, Gamma, Poisson, and Logistic (binary models only)). Information about the underlying record linkage process can be incorporated into the method if available (e.g., assumed overall mismatch rate, safe matches, predictors of match status, or predicted probabilities of correct matches).

License GPL-2

LazyData true

Encoding UTF-8

RoxygenNote 7.2.3

Maintainer Priyanjali Bukke <pbukke@gmu.edu>

URL https://github.com/bpriy/pldamixture

NeedsCompilation no

Author Priyanjali Bukke [aut, cre],
Zhenbang Wang [aut],
Martin Slawski [aut] (mslawsk3@gmu.edu),
Brady T. West [aut],
Emanuel Ben-David [aut],
Guoqing Diao [aut]

Repository CRAN

Date/Publication 2024-03-05 10:30:02 UTC

R topics documented:

`pldamixture-package` *Post-Linkage Data Analysis Based on Mixture Modelling*

Description

`pldamixture` implements the "General Framework for Regression with Mismatched Data" developed by Slawski et al., 2023. The framework uses a mixture model for pairs of linked records whose two components reflect distributions conditional on match status, i.e., correct match or mismatch. Inference is based on composite likelihood and the EM algorithm.

The package contains 4 functions for usage:
`fit_mixture`
`print.fitmixture`
`summary.fitmixture`
`predict.fitmixture`

Note

The references below discuss the implemented framework in more detail.

*Corresponding Author (mslawsk3@gmu.edu)

References

Slawski, M.*, West, B. T., Bukke, P., Diao, G., Wang, Z., & Ben-David, E. (2023). A General Framework for Regression with Mismatched Data Based on Mixture Modeling. Under Review. < doi: 10.48550/arXiv.2306.00909 >

Bukke, P., Ben-David, E., Diao, G., Slawski, M.*, & West, B. T. (2023). Cox Proportional Hazards Regression Using Linked Data: An Approach Based on Mixture Modelling. Under Review.

Slawski, M.*, Diao, G., Ben-David, E. (2021). A pseudo-likelihood approach to linear regression with partially shuffled data. Journal of Computational and Graphical Statistics. 30(4), 991-1003 < doi: 10.1080/10618600.2020.1870482 >

Examples

```
# optional inputs for linear regression of age at death on year of birth,
#    using a cubic polynomial specification.
## use commonness of names as predictors of match status
## first and last names were used for linkage
mformula <- ~commf + comml
## hand-linked records are considered "safe" matches
safematches <- ifelse(lifem$hndlnk =="Hand-Linked At Some Level", TRUE, FALSE)
## overall mismatch rate in the data set is assumed to be ~ 0.05
mrate <- 0.05

fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE), data = lifem,
                   family = "gaussian", mformula, safematches, mrate)
print(fit)
summary(fit)
predict(fit)
```

`fit_mixture` *Adjustment Method*

Description

Perform regression adjusted for mismatched data. The function currently supports Cox Proportional Hazards Regression (right-censored data only) and Generalized Linear Regression Models (Gaussian, Gamma, Poisson, and Logistic (binary models only)). Information about the underlying record linkage process can be incorporated into the method if available (e.g., assumed overall mismatch rate, safe matches, predictors of match status, or predicted probabilities of correct matches).

Usage

```
fit_mixture(
  formula,
  data,
  family = "gaussian",
  mformula,
  safematches,
  mrate,
 control = list(initbeta = "default", initgamma = "default", fy = "default", maxiter =
    1000, tol = 1e-04, cmaxiter = 1000),
  ...
)
```

Arguments

`formula`	a formula object for the outcome model, with the covariate(s) on the right of "~" and the response on the left. In the Cox proportional hazards setting, the response should be provided using the `Surv` function and the covariates should be separated by + signs.

data	a data.frame with linked data used in "formula" and "formula.m" (optional)
family	the type of regression model ("gaussian" - default, "poisson", "binomial", "gamma", "cox"). For Generalized Linear Models, standard link functions are used ("identity" for Gaussian, "log" for Poisson and Gamma, and "logit" for binomial).
mformula	a one-sided formula object for the mismatch indicator model, with the covariates on the right of "~". The default is an intercept-only model corresponding to a constant mismatch rate)
safematches	an indicator variable for safe matches (TRUE : record can be treated as a correct match and FALSE : record may be mismatched). The default is FALSE for all matches.
mrate	the assumed overall mismatch rate (a proportion between 0 and 1). If not provided, no overall mismatch rate is assumed.
control	an optional list variable to customize the initial parameter estimates ("initbeta" for the outcome model and "initgamma" for the mismatch indicator model), estimated marginal density of the response ("fy"), maximum iterations for the EM algorithm ("maxiter"), maximum iterations for the subroutine in the constrained logistic regression function ("cmaxiter"), and convergence tolerance for the termination of the EM algorithm ("tol").
...	the option to directly pass "control" arguments

Value

a list of results from the function called depending on the "family" specified.

coefficients	the outcome model coefficient estimates
match.prob	the correct match probabilities for all observations
objective	a variable that tracks the negative log pseudo-likelihood for all iterations of the EM algorithm.
family	the type of (outcome) regression model
standard.errors	the estimated standard errors
m.coefficients	the correct match model coefficient estimates
call	the matched call
wfit	an internal-use object for the predict function
dispersion	the dispersion parameter estimate when the family is a Generalized Linear Model
Lambdahat_0	the baseline cumulative hazard (using weighted Breslow estimator) when the family is "cox"
g_Lambdahat_0	the baseline cumulative hazard for the marginal density of the response variable (using Nelson-Aalen estimator) when the family is "cox"

Note

The references below discuss the implemented framework in more detail. The standard errors are estimated using Louis' method for the "cox" family (Bukke et al., 2023) and using the sandwich formula otherwise (Slawski et al., 2023).

*Corresponding Author (mslawsk3@gmu.edu)

References

Slawski, M.*, West, B. T., Bukke, P., Diao, G., Wang, Z., & Ben-David, E. (2023). A General Framework for Regression with Mismatched Data Based on Mixture Modeling. Under Review. < doi: 10.48550/arXiv.2306.00909 >

Bukke, P., Ben-David, E., Diao, G., Slawski, M.*, & West, B. T. (2023). Cox Proportional Hazards Regression Using Linked Data: An Approach Based on Mixture Modelling. Under Review.

Slawski, M.*, Diao, G., Ben-David, E. (2021). A pseudo-likelihood approach to linear regression with partially shuffled data. Journal of Computational and Graphical Statistics. 30(4), 991-1003 < doi: 10.1080/10618600.2020.1870482 >

Examples

```
## commonness score of first and last names used for linkage
mformula <- ~commf + comml
## hand-linked records are considered "safe" matches
safematches <- ifelse(lifem$hndlnk =="Hand-Linked At Some Level", TRUE, FALSE)
## overall mismatch rate in the data set is assumed to be ~ 0.05
mrate <- 0.05

fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE), data = lifem,
                   family = "gaussian", mformula, safematches, mrate)
```

`lifem` *LIFE-M Data*

Description

The `lifem` data set contains a subset of data from the Life-M project (https://life-m.org/) on 3,238 individuals born between 1883 to 1906. These records were obtained from linking birth certificates and death certificates either of two ways. A fraction of the records (2,159 records) were randomly sampled to be "hand-linked at some level" (HL). These records are high quality and were manually linked at some point by trained research assistants. The remaining records were "purely machine-linked" (ML) based on probabilistic record linkage without clerical review. The Life-M team expects the mismatch rate among these records to be around 5% (Bailey et al. 2022). Of interest is the relationship between age at death and year of birth. The `lifem` demo data set consists of 2,159 hand-linked records and 1,079 records that were randomly sampled from the purely machine-linked records (~2:1 HL-ML ratio).

Usage

```
data(lifem)
```

Format

a data frame with 3,238 rows and 6 variables

Details

- yob: year of birth (value from 1883 and 1906)
- unit_yob: yob re-scaled to the unit interval for analysis (between 0 and 1). If X is the yob, we use the following: (X – min(X)) / (max(X) – min(X)) = a * X + b, a = 1/(max(X) – min(X)), b = -min(X)*a
- age_at_death: age at death (in years)
- hndlnk: whether record was purely machine-linked or hand-linked at some level.
- commf: commonness score of first name (between 0 and 1). It is based on the 1940 census. It is a ratio of the log count of the individual's first name over the log count of the most commonly occurring first name in the census.
- comml: commonness score of last name (between 0 and 1). It is based on the 1940 census. It is a ratio of the log count of the individual's last name over the log count of the most commonly occurring last name in the census.

References

Bailey, Martha J., Lin, Peter Z., Mohammed, A.R. Shariq, Mohnen, Paul, Murray, Jared, Zhang, Mengying, and Prettyman, Alexa. LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (distributor), 2022-12-21. < doi: 10.3886/E155186V5 >

`predict.fitmixture` *Predictions From a "fitmixture" Object*

Description

Obtain predictions from a `fit_mixture()` object using `predict.coxph()`, `predict.glm()`, or `predict.lm()`.

Usage

```
## S3 method for class 'fitmixture'
predict(
  object,
  newdata,
  type,
  terms = NULL,
  na.action = na.pass,
  reference = "strata",
  ...
)
```

Arguments

object	the result of a call to fit_mixture()
newdata	optional new data to obtain predictions for. The original data is used by default.
type	the type of prediction. For the "cox" family, the choices are the linear predictor ("lp"), the risk score exp(lp) ("risk"), the expected number of events given the covariates and follow-up time ("expected"), and the terms of the linear predictor ("terms"). The survival probability for a subject is equal to exp(-expected). For the "gaussian" family, the choices are response ("response") or model term ("terms"). For the other glm families ("poisson", "binomial", "gamma"), the choices are predictions on the scale of the linear predictors ("link"), response ("response"), or model term ("terms").
terms	the terms when type = "terms". By default, all terms are included.
na.action	a function for what to do with missing values in newdata. The default is to predict "NA".
reference	when family = "cox", reference for centering predictions. Available options are c("strata" - default, "sample", "zero"). The default is "strata".
...	for future predict arguments

Value

a vector or matrix of predictions based on arguments specified.

Examples

```
## commonness score of first and last names used for linkage
mformula <- ~commf + comml
## hand-linked records are considered "safe" matches
safematches <- ifelse(lifem$hndlnk =="Hand-Linked At Some Level", TRUE, FALSE)
## overall mismatch rate in the data set is assumed to be ~ 0.05
mrate <- 0.05
fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE), data = lifem,
                   family = "gaussian", mformula, safematches, mrate)

predict(fit)
```

---

print.fitmixture *Print a "fitmixture" Object*

Description

Print call and outcome model coefficients from a fit_mixture() object

Usage

```
## S3 method for class 'fitmixture'
print(x, digits = max(3L, getOption("digits") - 3L), ...)
```

Arguments

x	the result of a call to fit_mixture()
digits	the number of significant digits to print
...	for additional print arguments

Value

invisibly returns the fit_mixture() object that is provided as an argument

Examples

```
## commonness score of first and last names used for linkage
mformula <- ~commf + comml
## hand-linked records are considered "safe" matches
safematches <- ifelse(lifem$hndlnk =="Hand-Linked At Some Level", TRUE, FALSE)
## overall mismatch rate in the data set is assumed to be ~ 0.05
mrate <- 0.05
fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE), data = lifem,
                   family = "gaussian", mformula, safematches, mrate)

print(fit)
```

---

summary.fitmixture *Summarize a "fitmixture" Object*

Description

Summarize results from a fit_mixture() object

Usage

```
## S3 method for class 'fitmixture'
summary(object, ...)
```

Arguments

object	the result of a call to fit_mixture()
...	for additional summary arguments

Value

a list of results from the function called depending on the "family" specified.

call	the matched call
family	the assumed type of (outcome) regression model

coefficients	a matrix with the outcome model's coefficient estimates, standard errors, t or z values, and p-values
m.coefficients	a matrix with the correct match model's coefficient estimates and standard errors
avgcmr	the average correct match rate among all records
match.prob	the correct match probabilities for all observations
dispersion	the dispersion parameter estimate when the family is a Generalized Linear Model

Examples

```
## commonness score of first and last names used for linkage
mformula <- ~commf + comml
## hand-linked records are considered "safe" matches
safematches <- ifelse(lifem$hndlnk =="Hand-Linked At Some Level", TRUE, FALSE)
## overall mismatch rate in the data set is assumed to be ~ 0.05
mrate <- 0.05
fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE), data = lifem,
                   family = "gaussian", mformula, safematches, mrate)

summary(fit)
```

Index