Package ‘FamEvent’

March 21, 2019

Type Package

Title Family Age-at-Onset Data Simulation and Penetrance Estimation

Version 2.0

Date 2019-03-12

Author Yun-Hee Choi, Karen Kopciuk, Wenqing He, Laurent Briollais

Maintainer Yun-Hee Choi <yun-hee.choi@schulich.uwo.ca>

Description Simulates age-at-onset traits associated with a segregating major gene in family data obtained from population-based, clinic-based, or multi-stage designs. Appropriate ascertainment correction is utilized to estimate age-dependent penetrance functions either parametrically from the fitted model or nonparametrically from the data. The Expectation and Maximization algorithm can infer missing genotypes and carrier probabilities estimated from family's genotype and phenotype information or from a fitted model. Plot functions include pedigrees of simulated families and predicted penetrance curves based on specified parameter values.

License GPL (>= 2.0)

NeedsCompilation no

LazyData true

Depends R (>= 3.0.0), survival

Imports MASS, kinship2, truncnorm, eha, pracma

Repository CRAN

Date/Publication 2019-03-21 05:13:22 UTC

R topics documented:

FamEvent-package *Family age-at-onset data simulation and penetrance estimation*

Description

Family-based studies are used to characterize the disease risk associated with being a carrier of a major gene. When the disease risk can vary with age of onset, penetrance or disease risk functions need to provide age-dependent estimates of this disease risk over lifetime. This FamEvent package can generate age-at-onset data in the context of familial studies, with correction for ascertainment (selection) bias arising from a specified study design based on proband's mutation and disease statuses. Possible study designs are: "pop" for population-based design where families are ascertained through affected probands, "pop+" are similar to "pop" but probands are also known mutation carriers, "cli" for clinic-based design that includes affected probands with at least one parent and one sib affected, "cli+" are similar to "cli" but probands are also known mutation carrriers. And "twostage" for two-stage design that randomly samples families from the population in the first stage and oversamples high risk families that includes at least two affected members in the family at the second stage.

Ages at disease onset are generated specific to family members' gender and mutation status according to the specified study design with residual familial correlations induced by either a shared frailty or a second gene. For estimating age at onset risks with family data, an ascertainment corrected prospective likelihood approach is used to account for the population or clinic-based study designs while a composite likelihood approach is used for the two-stage sampling design. The Expectation and Maximization (EM) algorithm has been implemented for inferring missing genotypes conditional on observed genotypes and phenotypes in the families. For family members who have missing genotypes, their carrier probabilities are obtained either from the fitted model or from Mendelian transmission probabilities. This package also provides functions to plot the age-dependent penetrance curves estimated parametrically from the fitted model or non-parametrically from the data, pedigree plots of simulated families and penetrance function curves for carriers and non-carriers of a major and second gene based on specified parameter values.

Author(s)

Yun-Hee Choi, Karen Kopciuk, Laurent Briollais, Wenqing He

Maintainer: Yun-Hee Choi < yun-hee.choi@schulich.uwo.ca >

References

Choi, Y.-H., Kopciuk, K. and Briollais, L. (2008) Estimating Disease Risk Associated Mutated Genes in Family-Based Designs, Human Heredity 66, 238-251

Choi, Y.-H. and Briollais (2011) An EM Composite Likelihood Approach for Multistage Sampling of Family Data with Missing Genetic Covariates, Statistica Sinica 21, 231-253

See Also

simfam, summary.simfam, plot.simfam, penplot, carrierprob, penmodel, penmodelEM, print.penmodel, summary.penmodel,print.summary.penmodel, plot.penmodel

Examples

```
#  Simulate family data

set.seed(4321)
fam <- simfam(N.fam = 100, design = "pop+", variation = "none", base.dist = "Weibull",
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35), allelefreq = 0.02)

# summary of simulated family data

summary(fam)

# Pedigree plots for family 1 and 2

plot(fam, famid = c(1,2))

# penetrance function plots given model parameter values for Weibull baseline

penplot(base.parms = c(0.01, 3), vbeta = c(-1.3, 2.35), base.dist = "Weibull",
        variation = "none", agemin = 20)

# model fit of family data

fit <- penmodel(Surv(time, status) ~ gender + mgene, cluster = "famID", design = "pop+",
       parms=c(0.01, 3, -1.13, 2.35), data = fam, base.dist = "Weibull", robust = TRUE)


# summary of estimated model parameters and penetrance estimates

summary(fit)

# penetrance curves useful for model checking

plot(fit)
```

`carrierprob` *Compute mutation carrier probabilities for individuals with missing gentoypes*

Description

Computes model- or data-based carrier probabilities for individuals with missing genotypes based on the observed mutation status of family members and the individual's phenotype.

Usage

```
carrierprob(condition = "geno", method = "data", fit = NULL, data, mode = "dominant",
q = 0.02)
```

Arguments

`condition`	Choice of conditional information to use for computing the carrier probability. Possible choices are `"geno"` for using observed genotypes and `"geno+pheno"` for using both observed genotype and phenotype information in the calculation of the carrier probability.
`method`	Choice of methods to calculate the carrier probability. Possible choices are `"data"` for empirical calculation of the carrier probabilities based on data, `"mendelian"` using Mendelian transmission probabilities based on observed carriers within families, or `"model"` using the parametric model fit; see details below. Default is `"data"`. If `method = "data"`, only `data` is required to be specified.
`fit`	An object of class `penmodel`, a fitted model by `penmodelEM` function for inferring missing mutation statuses in the family.
`data`	Family data that includes missing genotypes using the same data format generated by the function `simfam`.
`mode`	Choice of modes of inheritance when using `method="model"`. Possible choices are `"dominant"` for dominant model or `"recessive"` for recessive model. Default is `"dominant"`.
`q`	Frequency of the disease causing allele when using `method="model"`. The value should be between 0 and 1. If `NULL`, the estimated allele frequency from data will be used. Default value is `0.02`.

Details

When `method="model"` along with the choice of `condition="geno+pheno"`, the carrier probability for individual i is calculated by conditioning on her/his observed phenotype and carrier statuses of family members

$$P(X_i = 1|Y_i, X^o) = \frac{P(Y_i|X_i = 1)P(X_i = 1|X^o)}{P(Y_i|X_i = 1)P(X_i = 1|X^o) + P(Y_i|X_i = 0)P(X_i = 0|X^o)},$$

where X_i indicates the unknown carrier status of individual i and X^o represents the observed carrier statuses in his or her family members; Y_i represents the observed phenotype (t_i, δ_i) of individual i in terms of age at onset t_i and disease status indicator δ_i with 1 used for affected individuals and 0 for unaffected individuals.

When `method="mendelian"` along with the choice of `condition="geno"`, the carrier probability is calculated based on Mendelian laws of genetic transmission with a fixed allele frequency.

Value

Returns a data frame with a vector of carrier probabilities called `carrp.geno` when `condition="geno"` or `carrp.pheno` when `condtion="geno+pheno"` added after the last column of the family data.

Author(s)

Yun-Hee Choi

See Also

`simfam`, `penmodelEM`, `plot.simfam`, `summary.simfam`

Examples

```
# Simulated family data with 30% of members missing their genetic information.

set.seed(4321)
fam <- simfam(N.fam = 100, design = "pop+", base.dist = "Weibull", mrate = 0.3,
       base.parms = c(0.01,3), vbeta = c(-1.13, 2.35), agemin = 20)

# EM algorithm for fitting family data with missing genotypes assuming a Weibull
# baseline hazard and dominant mode of Mendelian inheritance for a major gene.

fitEM <- penmodelEM(Surv(time, status) ~ gender + mgene, cluster = "famID", gvar = "mgene",
      parms = c(0.01, 3, -1.13, 2.35), data = fam, design = "pop+", base.dist = "Weibull",
        method = "mendelian", mode = "dominant")

# Carrier probability obtained by conditioning on the observed genotypes and phenotype,
# assuming a dominant Mendelian mode of inheritance

fam.added <- carrierprob(condition = "geno+pheno", method = "model", fit = fitEM,
             data = fam, mode = "dominant", q = 0.02)

# pedigree plot for family 1 displaying carrier probabilities

plot.simfam(fam.added, famid = 1)
```

fampower *Simulation-based power calculation for genetic effect*

Description

Computes the power of detecting genetic effect in the penetrance model based on a family-based simulation study.

Usage

```
fampower(N.fam, N.sim, effectsize, beta.sex, alpha = 0.05, side = 2, design = "pop",
variation = "none", interaction = FALSE, depend = NULL, base.dist = "Weibull",
frailty.dist = NULL, base.parms, allelefreq = c(0.02, 0.2), dominant.m = TRUE,
dominant.s = TRUE, mrate = 0, hr = 0, probandage = c(45, 2), agemin = 20, agemax = 100)
```

Arguments

`N.fam`	Number of families to generate.
`N.sim`	Number of simulations.
`effectsize`	Effect size of the major mutated gene (`beta.gene`) to detect under the alternative hypothesis. When `interaction=TRUE`, both the main and interaction effects should be specified, `effectsize = c(beta.gene, beta.int)`.
`beta.sex`	Gender effect that is fixed in the model.
`alpha`	Significance level. Default value is 0.05.
`side`	Number of sides for the alternative hypothesis. Possible choices are 1 for one-sided test and 2 for two-sided test. Default value is 2.
`design`	Family based study design used in the simulations. Possible choices are: `"pop"`, `"pop+"`, `"cli"`, `"cli+"` or `"twostage"`, where `"pop"` is for the population-based design that families are ascertained by affected probands, `"pop+"` is similar to `"pop"` but with mutation carrier probands, `"cli"` is for the clinic-based design that includes affected probands with at least one parent and one sibling affected, `"cli+"` is similar to `"cli"` but with mutation carrier probands and `"twostage"` for two-stage design that randomly samples families from the population in the first stage and oversamples high risk families in the second stage that includes at least two affected members in the family. Default is `"pop"`.
`variation`	Source of residual familial correlation. Possible choices are: `"frailty"` for frailty shared within families, `"secondgene"` for second gene variation, or `"none"` for no residual familial correlation. Default is `"none"`.
`interaction`	Logical; if `TRUE`, the interaction between gender and mutation status is allowed, otherwise no interaction is allowed. Default is `FALSE`.
`depend`	Variance of the frailty distribution. Dependence within families increases with depend value. Default value is `NULL`. Value `> 0` should be specified when `variation = "frailty"`.

`base.dist` Choice of baseline hazard distribution. Possible choices are: `"Weibull"`, `"loglogistic"`, `"Gompertz"`, `"lognormal"` `"gamma"`, or `"logBurr"`. Default is `"Weibull"`.

`frailty.dist` Choice of frailty distribution. Possible choices are: `"gamma"` for gamma distribution or `"lognormal"` for log normal distribution when `variation = "frailty"`. Default is `NULL`.

`base.parms` Vector of parameter values for baseline hazard function.

`base.parms = c(lambda, rho)`, where lambda and rho are the shape and scale parameters, respectively. If `base.dist = "logBurr"` is chosen, three parameters should be specified for `base.parms = c(lambda, rho, eta)`.

`allelefreq` Vector of population allele frequencies of major and second disease gene alleles. Frequencies must be between 0 and 1. Default frequencies are 0.02 for major gene allele and 0.2 for second gene allele, `allelefreq=c(0.02, 0.2)`.

`dominant.m` Logical; if `TRUE`, the genetic model of the major gene is dominant, otherwise recessive.

`dominant.s` Logical; if `TRUE`, the genetic model of the second gene is dominant, otherwise recessive.

`mrate` Proportion of missing genotypes, value between 0 and 1. Default value is 0.

`hr` Proportion of high risk families, which include at least two affected members, to be sampled from the two stage sampling. This value should be specified when `design = "twostage"`. Default value is 0. Value should lie between 0 and 1.

`probandage` Vector of mean and standard deviation for the proband age. Default values are mean of 45 years and standard deviation of 2 years, `probandage = c(45, 2)`.

`agemin` Minimum age of disease onset or minimum age. Default is 20 years of age.

`agemax` Maximum age of disease onset or maximum age. Default is 100 years of age.

Details

The power of testing $H_0 : \beta_{gene} = 0$ vs. $H_1 : \beta_{gene} =$ `effectsize` is obtained by the proportion of times the null hypothesis is rejected out of the `N.sim` simulations.

When `interaction = TRUE`, the powers of both the main effect of mutated gend and the interaction effect of mutated gene and gender will be computed.

Value

Returns

`power` Power of detecting the genetic effect.

Author(s)

Yun-Hee Choi

See Also

`simfam`

Examples

```
## Example 1: obtain the power for testing the genetic effect
# based on 50 POP families simulated using 100 simulations
## Not run:
set.seed(4321)
fampower(N.fam = 50, N.sim = 100, effectsize = 1, beta.sex = 0.8, alpha = 0.05, side = 2,
design = "pop+", variation = "none", base.dist = "Weibull", allelefreq = 0.02,
base.parms = c(0.01, 3))
## End(Not run)

## Example 2: obtain the power for both the main and interaction effects
# based on 50 POP families simulated using 100 simulations
## Not run:
set.seed(4321)
fampower(N.fam = 50, N.sim = 100, effectsize = c(1.5, 1), beta.sex = 0.8, alpha = 0.05,
side = 2, interaction = TRUE,  design = "pop+", variation = "none", base.dist = "Weibull",
allelefreq = 0.02, base.parms = c(0.01, 3))
## End(Not run)
```

LSfam *Ontario Lynch Syndrom families*

Description

Data from 32 Lynch Syndrome families segregating mismatch repair mutations selected from the Ontario Familial Colorectal Cancer Registry that includes 765 individuals, both probands and relatives. The families were ascertained throughout affected and mutation carrier probands.

Usage

```
data("LSfam")
```

Format

A data frame with 765 observations on the following 11 variables.

`famID` Family identification (ID) numbers.

`indID` Individuals ID numbers.

`fatherID` Father ID numbers.

`motherID` Mother ID numbers.

`gender` Gender indicators: 1 for male, 0 for female.

`status` Disease statuses: 1 for affected, 0 for unaffected.

`time` Ages at diagnosis of colorectal cancer for the affected or ages of last follow-up for the unaffected.

`currentage` Current ages in years.

`mgene` MLH1 or MSH2 mutation indicators: 1 for mutated gene carriers, 0 for mutated gene non-carriers, or `NA` if missing.

`proband` Proband indicators: 1 for proband, 0 for non-proband.

`relation` Family members' relationship with the proband.

1	Proband (self)
2	Brother or sister
3	Son or daughter
4	Parent
5	Nephew or niece
6	Spouse
7	Brother or sister in law
8	Paternal grandparent
9	Paternal uncle or aunt
10	Paternal cousin
11	Maternal grandparent
12	Maternal uncle or aunt
13	Maternal cousin
14	Son or daughter in law
15	Grandchild
16	Uncle's or aunt's spouse.

References

Choi, Y.-H., Cotterchio, M., McKeown-Eyssen, G., Neerav, M., Bapat, B., Boyd, K., Gallinger, S., McLaughlin, J., Aronson, M., and Briollais, L. (2009). Penetrance of Colorectal Cancer among MLH1/ MSH2 Carriers Participating in the Colorectal Cancer Familial Registry in Ontario, Hereditary Cancer in Clinical Practice, 7:14.

Examples

```
data(LSfam)

# Summary of LSfam
summary.simfam(LSfam)

# Pedigree plot for the first family
plot.simfam(LSfam)

# Assign minimum age for fitting penmodel
attr(LSfam, "agemin") <- 18

fit <- penmodelEM(Surv(time, status) ~ gender + mgene, cluster = "famID",
       parms = c(0.05, 2, 1, 3), data = LSfam[!is.na(LSfam$time) & LSfam$time > 18, ],
       method = "mendelian", base.dist = "Weibull", design = "pop+", robust = TRUE)

summary(fit)
```

```
penetrance(fit, fixed = c(1, 1), age = c(50, 60, 70), CI = TRUE, MC = 100)
```

penetrance *Penetrance function and confidence intervals*

Description

Estimates the cumulative disease risks (penetrances) and confidence intervals at given age(s) based on the fitted penetrance model.

Usage

```
penetrance(fit, fixed, age, CI = TRUE, MC = 100)
```

Arguments

`fit`	An object class of `'penmodel'`, a fitted model by `penmodel` or `penmodelEM` functions.
`fixed`	Vector of fixed values of the covariates used for penetrance calculation.
`age`	Vector of ages used for penetrance calculation.
`CI`	Logical; if `TRUE`, the 95% confidence interval will be obtained using a Monte-Carlo method, otherwise no confidence interval will be provided. Default is `TRUE`.
`MC`	Number of simulated samples used to calculate confidence intervals with a Monte-Carlo method. If `MC=0`, no confidence intervals will be calculated. Default value is 100.

Details

The penetrance function is defined as the probability of developing a disease by age t given fixed values of covariates x,

$$P(T < t|x) = 1 - S(t; x),$$

where t is greater than the minimum age t_0 and $S(t; x)$ is the survival distribution based on a proportional hazards model with a specified baseline hazard distribution.

The proportional hazards model is specified as:

$$h(t|x) = h_0(t) \exp(\beta x),$$

where $h_0(t)$ is the baseline hazards function, x is the vector of covariates and β is the vector of corresponding regression coefficients.

Calculations of standard errors of the penetrance estimates and 95% confidence intervals (CIs) for the penetrance at a given age are based on Monte-Carlo simulations of the estimated penetrance model.

A multivariate normal distribution is assumed for the parameter estimates, and `MC = n` sets of the parameters are generated from the multivariate normal distribution with the parameter estimates and their variance-covariance matrix. For each simulated set, a penetrance estimate is calculated at a given age by substituting the simulated parameters into the penetrance function.

The standard error of the penetrance estimate at a given age is calculated by the standard deviation of penetrance estimates obtained from n simulations.

The 95% CI for the penetrance at a given age is calculated using the 2.5th and 97.5th percentiles of the penetrance estimates obtained from n simulations.

Value

Returns the following values:

`age` Ages at which the penetrances are calculated.

`penetrance` Penetrance estimates at given ages.

`lower` Lower limit of the 95% confidence interval; simulation-based 2.5th percentile of the penetrance estimates.

`upper` Upper limit of the 95% confidence interval; simulation-based 97.5th percentile of the penetrance estimates.

`se` Simulation-based standard errors of the penetrance estimates.

Author(s)

Yun-Hee Choi

See Also

`simfam`, `penmodel`, `penmodelEM`

Examples

```
 set.seed(4321)
 fam <- simfam(N.fam = 100, design = "pop+", base.dist = "Weibull", allelefreq = 0.02,
        base.parms = c(0.01,3), vbeta = c(-1.13, 2.35))

 fit <- penmodel(Surv(time, status) ~ gender +  mgene, cluster = "famID",
      parms = c(0.01, 3, -1.13, 2.35),  data = fam, base.dist = "Weibull", design = "pop+")

# Compute penetrance estimates for male carriers at age 40, 50, 60, and 70 and
# their 95% CIs based on 100 Monte Carlo simulations.

penetrance(fit, fixed = c(1,1), age = c(40, 50, 60, 70), CI = TRUE, MC = 100)
```

penmodel *Fit a penetrance model*

Description

Fits a penetrance model for family data based on a prospective likelihood with ascertainment correction and provides model parameter estimates.

Usage

```
penmodel(formula, cluster = "famID", gvar = "mgene", parms, cuts = NULL, data,
design = "pop", base.dist = "Weibull", agemin = NULL, robust = FALSE)
```

Arguments

`formula` A formula expression as for other regression models. The response should be a survival object as returned by the `Surv` function. See the documentation for `Surv`, `lm` and `formula` for details.

`cluster` Name of cluster variable. Default is `"famID"`.

`gvar` Name of genetic variable. Default is `"mgene"`.

`parms` Vector of initial values for the parameters in the model including baseline parameters and regression coefficients. `parms = c(baseparm, coef)`, where `baseparm` includes the initial values for baseline parameters used for `base.dist`, and `coef` includes the initial values for regression coefficients for the variables specified in `formula`. See details for the baseline parameters.

`cuts` Vector of cut points that define the intervals where the hazard function is constant. The `cuts` should be specified when `base.dist="piecewise"` and must be strictly positive and finite. Default is `NULL`.

`data` Data frame generated from `simfam` or data frame containing variables named in the `formula` and specific variables: `famID`, `indID`, `gender`, `currentage`, `mgene`, `time`, `status` and `weight` with `attr(data,"agemin")` specified.

`design` Study design of the family data. Possible choices are: `"pop"`, `"pop+"`, `"cli"`, `"cli+"` or `"twostage"`, where `"pop"` is for the population-based design with affected probands whose mutation status can be either carrier or non-carrier, `"pop+"` is similar to `"pop"` but with mutation carrier probands, `"cli"` is for the clinic-based design that includes affected probands with at least one parent and one sib affected, `"cli+"` is similar to `"cli"` but with mutation carrier probands, and `"twostage"` is for the two-stage design with oversampling of high risks families. Default is `"pop"`.

`base.dist` Choice of baseline hazard distributions to fit. Possible choices are: `"Weibull"`, `"loglogistic"`, `"Gompertz"`, `"lognormal"`, `"gamma"`, `"logBurr"`, or `"piecewise"`. Default is `"Weibull"`.

`agemin` Minimum age of disease onset or minimum age. Default is `NULL`.

`robust` Logical; if `TRUE`, the robust ‘sandwich’ standard errors and variance-covariance matrix are provided, otherwise the conventional standard errors and variance-covariance matrix are provided.

Details

The penetrance model is fitted to family data with a specified baseline hazard distribution,

$$h(t|x_s, x_g) = h_0(t - t_0) \exp(\beta_s x_s + \beta_g x_g),$$

where $h_0(t)$ is the baseline hazards function specified by `base.dist`, which depends on the shape and scale parameters, λ and ρ; x_s indicates male (1) and female (0) and x_g indicates carrier (1) or non-carrier (0) of a gene of interest (major gene). Additional covariates can be added to `formula` in the model.

For family data arising from population- or clinic-based study designs (`design="pop", "pop+"`, `"cli"`, or `"cli+"`), the parameters of the penetrance model are estimated using the ascertainment-corrected prospective likelihood approach (Choi, Kopciuk and Briollais, 2008).

For family data arising from a two-stage study design (`design="twostage"`), model parameters are estimated using the composite likelihood approach (Choi and Briollais, 2011)

Note that the baseline parameters include `lambda` and `rho`, which represent the scale and shape parameters, respectively, and `eta`, additional parameter to specify for `"logBurr"` distribution. For the `"lognormal"` baseline distribution, `lambda` and `rho` represent the location and scale parameters for the normally distributed logarithm, where `lambda` can take any real values and `rho` > 0. For the other baselinse distributions, `lambda` > 0, `rho` > 0, and `eta` > 0. When a piecewise constant distribution is specified for the baseline hazards, `base.dist="piecewise"`, `baseparm` should specify the initial interval-constant values, one more than the cut points specified by`cuts`.

Transformed baseline parameters are used for estimation; log transformation is applied to both scale and shape parameters (λ, ρ) for `"Weibull"`, `"loglogistic"`, `"Gompertz"` and `"gamma"` baselines, to (λ, ρ, η) for `"logBurr"` and to the piecewise constant parameters for a `piecewise` baseline hazard. For `"lognormal"` baseline distribution, the log transformation is applied only to ρ, not to λ, which represents the location parameter for the normally distributed logarithm.

Calculations of penetrance estimates and their standard errors and 95% confidence intervals at given ages can be obtained by `penetrance` function via Monte-Carlo simulations of the estimated penetrance model.

Value

Returns an object of class `'penmodel'`, including the following elements:

`estimates` Parameter estimates of transformed baseline parameters and regression coefficients.

`varcov` Variance-covariance matrix of parameter estimates obtained from the inverse of Hessian matrix.

`varcov.robust` Robust ‘sandwich’ variance-covariance matrix of parameter estimates when `robust=TRUE`.

`se` Standard errors of parameter estimates obtained from the inverse of Hessian matrix.

`se.robust` Robust ‘sandwich’ standard errors of parameter estimates when `robust=TRUE`.

`logLik` Loglikelihood value for the fitted penetrance model.

`AIC` Akaike information criterion (AIC) value of the model; AIC = 2*k - 2*logLik, where k is the number of parameters used in the model.

Author(s)

Yun-Hee Choi

References

Choi, Y.-H., Kopciuk, K. and Briollais, L. (2008) Estimating Disease Risk Associated Mutated Genes in Family-Based Designs, Human Heredity 66, 238-251

Choi, Y.-H. and Briollais (2011) An EM Composite Likelihood Approach for Multistage Sampling of Family Data with Missing Genetic Covariates, Statistica Sinica 21, 231-253

See Also

penmodelEM, simfam, penplot, print.penmodel, summary.penmodel, print.summary.penmodel, plot.penmodel

Examples

```
# Family data simulated from population-based design using a Weibull baseline hazard

set.seed(4321)
fam <- simfam(N.fam = 200, design = "pop+", variation = "none", base.dist = "Weibull",
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35), agemin = 20, allelefreq = 0.02)

# Penetrance model fit for simulated family data

fit <- penmodel(Surv(time, status) ~ gender + mgene, cluster = "famID", design = "pop+",
       parms = c(0.01, 3, -1.13, 2.35), data = fam, base.dist = "Weibull")

# Summary of the model parameter estimates from the model fit

summary(fit)

# Plot the lifetime penetrance curves with 95% CIs from the model fit for specific
# gender and mutation status groups along with their nonparametric penetrance curves
# based on data excluding probands.

plot(fit, add.KM = TRUE, conf.int = TRUE, MC = 100)
```

---

penmodelEM *EM algorithm for estimating the penetrance model with missing genotypes*

Description

Fits a penetrance model for family data with missing genotypes via the EM algorithm and provides model parameter estimates.

Usage

```
penmodelEM(formula, cluster = "famID", gvar = "mgene", parms, cuts = NULL, data,
design = "pop", base.dist = "Weibull", agemin = NULL, robust = FALSE, method = "data",
mode = "dominant", q = 0.02)
```

Arguments

formula
A formula expression as for other regression models. The response should be a survival object as returned by the Surv function. See the documentation for Surv, lm and formula for details.

cluster
Name of cluster variable. Default is "famID".

gvar
Name of genetic variable. Default is "mgene".

parms
Vector of initial values for the parameters in the model including baseline parameters and regression coefficients. parms = c(baseparm, coef), where baseparm includes the initial values for baseline parameters used for base.dist, and coef includes the initial values for regression coefficients for the variables specified in formula. See details for the baseline parameters.

cuts
Vector of cuts that define the intervals where the hazard function is constant. The cuts should be specified base.dist="piecewise" and must be strictly positive and finite. Default is NULL.

data
Data frame generated from simfam or data frame containing specific variables: famID, indID, gender, currentage, mgene, time, status and weight with attr(data,"agemin") specified.

design
Study design of the family data. Possible choices are: "pop", "pop+", "cli", "cli+" or "twostage", where "pop" is for the population-based design with affected probands whose mutation status can be either carrier or non-carrier, "pop+" is similar to "pop" but with mutation carrier probands, "cli" is for the clinic-based design that includes affected probands with at least one parent and one sib affected, "cli+" is similar to "cli" but with mutation carrier probands, and "twostage" is for the two-stage design with oversampling of high risks families. Default is "pop".

base.dist
Choice of baseline hazard distributions to fit. Possible choices are: "Weibull", "loglogistic", "Gompertz", "lognormal", "gamma", "logBurr", or "piecewise". Default is "Weibull".

agemin
Minimum age of disease onset or minimum age. Default is NULL.

robust
Logical; if TRUE, the robust ‘sandwich’ standard errors and variance-covariance matrix are provided, otherwise the conventional standard errors and variance-covariance matrix are provided.

method
Choice of methods for calculating the carrier probabilities for individuals with missing mutation status. Possible choices are "data" for empirical calculation of the carrier probabilities based on the observed carriers’ statuses in the entire sample, specific to generation and proband’s mutation status or "mendelian" for calculating carrier probabilities based on Mendelian transmission probabilities with the given allele frequency and mutation statuses observed in the family. Default is "data".

If `method = "mendelian"`, specify both `mode` of the inheritance and the allele frequency `q`.

`mode` Choice of modes of inheritance for calculating carrier probabilies for individuals with missing mutation status. Possible choices are `"dominant"` for dominant model or `"recessive"` for recessive model. Default is `"dominant"`.

`q` Frequency of the disease causing allele used for calculating carrier pobabilities. The value should be between 0 and 1. If `NULL`, the estimated allele frequency from data will be used. Default value is `0.02`.

Details

The expectation and maximization (EM) algorithm is applied for making inference about the missing genotypes. In the expectation step, for individuals with unknown carrier status, we first compute their carrier probabilities given their family's observed phenotype and genotype information based on current estimates of parameters θ as follows,

$$w_{fi} = P(X_{fi} = 1 | Y_{fi}, X_f^o),$$

where X_{fi} represents the mutation carrier status and Y_{fi} represents the phenotype in terms of age at onset t_{fi} and disease status δ_{fi} for individual $i, i = 1, \ldots, n_f$, in family $f, f = 1, ..., n$, and X_f^o represents the observed genotypes in family f.

Then, we obtain the conditional expectation of the log-likelihood function (ℓ) of the complete data given the observed data as a weighted log-likelihood, which has the form

$$E_\theta[\ell(\theta|Y, X^o)] = \sum_f^n \sum_i^{n_f} \ell_{fi}(\theta|X_{fi} = 1) w_{fi} + \ell_{fi}(\theta|X_{fi} = 0)(1 - w_{fi}).$$

In the maximization step, the updated parameter estimates are obtained by maximizing the weighted log likelihood computed in the E-step. These expectation and maximization steps iterate until convergence to obtain the maximum likelihood estimates. See more details in Choi and Briollais (2011) or Choi et al. (2014).

Note that the baseline parameters include `lambda` and `rho`, which represent the scale and shape parameters, respectively, and `eta`, additional parameter to specify for `"logBurr"` distribution. For the `"lognormal"` baseline distribution, `lambda` and `rho` represent the location and scale parameters for the normally distributed logarithm, where `lambda` can take any real values and `rho` > 0. For the other baselinse distributions, `lambda` > 0, `rho` > 0, and `eta` > 0. When a piecewise constant distribution is specified for the baseline hazards, `base.dist="piecewise"`, `baseparm` should specify the initial interval-constant values, one more than the cut points specified by`cuts`.

Transformed baseline parameters are used for estimation; log transformation is applied to both scale and shape parameters (λ, ρ) for `"Weibull"`, `"loglogistic"`, `"Gompertz"` and `"gamma"` baselines, to (λ, ρ, η) for `"logBurr"` and to the piecewise constant parameters for a `piecewise` baseline hazard. For `"lognormal"` baseline distribution, the log transformation is applied only to ρ, not to λ, which represents the location parameter for the normally distributed logarithm.

Calculations of penetrance estimates and their standard errors and 95% confidence intervals at given ages can be obtained by `penetrance` function via Monte-Carlo simulations of the estimated penetrance model.

Value

Returns an object of class 'penmodel', including the following elements:

`estimates` Parameter estimates of transformed baseline parameters and regression coefficients.

`varcov` Variance-covariance matrix of parameter estimates obtained from the inverse of Hessian matrix.

`varcov.robust` Robust ‘sandwich’ variance-covariance matrix of parameter estimates when `robust=TRUE`.

`se` Standard errors of parameter estimates obtained from the inverse of Hessian matrix.

`se.robust` Robust ‘sandwich’ standard errors of parameter estimates when `robust=TRUE`.

`logLik` Loglikelihood value for the fitted penetrance model.

`AIC` Akaike information criterion (AIC) value of the model; AIC = 2*k - 2*logLik, where k is the number of parameters used in the model.

Author(s)

Yun-Hee Choi

References

Choi, Y.-H. and Briollais, L. (2011) An EM composite likelihood approach for multistage sampling of family data with missing genetic covariates, Statistica Sinica 21, 231-253.

Choi, Y.-H., Briollais, L., Green, J., Parfrey, P., and Kopciuk, K. (2014) Estimating successive cancer risks in Lynch Syndrome families using a progressive three-state model, Statistics in Medicine 33, 618-638.

See Also

`simfam`, `penmodel`, `print.penmodel`, `summary.penmodel`, `print.summary.penmodel`, `plot.penmodel`, `carrierprob`

Examples

```
# Family data simulated with 20% of members missing their genetic information.

set.seed(4321)
fam <- simfam(N.fam = 100, design = "pop+", base.dist = "Weibull", base.parms = c(0.01, 3),
       vbeta = c(1, 2), agemin = 20, allelefreq = 0.02, mrate = 0.2)

# EM algorithm for fitting family data with missing genotypes

fit <- penmodelEM(Surv(time, status) ~ gender + mgene, cluster = "famID", gvar = "mgene",
       parms = c(0.01, 3, 1, 2), data = fam, design="pop+", robust = TRUE,
       base.dist = "Weibull", method = "mendelian", mode = "dominant", q = 0.02)

# Summary of the model parameter estimates from the model fit by penmodelEM
```

```
summary(fit)

# Plot the lifetime penetrance curves from model fit for gender and
# mutation status groups along with their nonparametric penetrance curves
# based on observed data excluding probands.

plot(fit)
```

penplot *Plot penetrance functions*

Description

Plots the penetrance functions given the values of baseline parameters and regression coefficients and choices of baseline and frailty distributions.

Usage

```
penplot(base.parms, vbeta, cuts = NULL, variation = "none", base.dist = "Weibull",
frailty.dist = NULL, depend = 1, agemin = 20, agemax = 80, print = TRUE,
col = c("blue","red","blue","red"),  lty = c(1, 1, 2, 2), add.legend = TRUE,
add.title = TRUE, x = "topleft", y = NULL, xlab = "Age at onset", ylab = "Penetrance",
ylim = NULL, main = NULL, ...)
```

Arguments

`base.parms` Vector of parameter values for the specified baseline hazard function: `base.parms = c(lambda, rho)` should be specified for `base.dist = "Weibull"`, `"loglogistic"`, `"Gompertz"`, `"gamma"`, and `"lognormal"`, `c(lambda, rho, eta)` for `base.dist = "logBurr"`, or interval constant hazard values for the intervals produced by `cuts` for `base.dist = "piecewise"`.

`vbeta` Vector of regression coefficients for gender and majorgene, `vbeta = c(beta.s, beta.g)`. If `variation = "secondgene"`, regression coefficients for gender, major gene and second gene, `vbeta = c(beta.s, beta.g1, beta.g2)`, should be specified.

`cuts` Vector of cut points defining the intervals where the hazard function is constant. The `cuts` should be specified when `base.dist = "piecewise"` and must be strictly positive and finite. Default is `NULL`.

`variation` Source of residual familial correlation. Possible choices are: `"frailty"` for frailty shared within families, `"secondgene"` for second gene variation, or `"none"` for no residual familial correlation. Default is `"none"`.

`base.dist` Choice of baseline hazard distribution. Possible choices are: `"Weibull"`, `"loglogistic"`, `"Gompertz"`, `"lognormal"`, `"gamma"`, or `"piecewise"`. Default is `"Weibull"`.

`frailty.dist` Choice of frailty distribution. Possible choices are `"gamma"` for gamma distribution or `"lognormal"` for log normal distributions when `variation = "frailty"`. Default is `NULL`.

`depend` Variance of the frailty distribution. Dependence within families increases with `depend` value. Default value is 1.

`agemin` Minimum age of disease onset. Default is 20 years of age.

`agemax` Maximum age of disease onset. Default is 80 years of age.

`print` Logical; if `TRUE`, prints the penetrance values by age 70 obtained from the assumed model for male carriers, female carriers, male noncarrers, and female noncarriers. Default is `TRUE`.

`col` Colors of lines for male carriers, female carriers, male noncarrers, and female noncarriers. Default is `c("blue", "red", "blue", "red")`.

`lty` Types of lines for male carriers, female carriers, male noncarriers, and female noncarriers. Default is `c(1, 1, 2, 2)`.

`add.legend` Logical; if `TRUE`, displays legend in the plot. Default is `TRUE`.

`add.title` Logical; if `TRUE`, displays title in the plot. Default is `TRUE`.

`x, y` Position of legend; see legend. Defaults are `x = "topleft", y = NULL`.

`xlab` Title for the x-axis. Default is `"Age at onset"`.

`ylab` Title for the y-axis. Default is `"Penetrance"`.

`ylim` Limits of the y-axis. Default is `NULL`. If `NULL`, `ylim` will be automatically determined.

`main` Main title of the plot. Default is `NULL`. If `NULL`, the title will be automatically created.

`...` Other parameters to be passed through to plotting functions.

Details

Proportional hazard models The penetrance model conditional on the covariates $X = (x_s, x_g)$ is assumed to have the following hazard function:

$$h(t|X) = h_0(t - t_0)\exp(\beta_s x_s + \beta_g x_g),$$

where $h_0(t)$ is the baseline hazard function, t_0 is a minimum age of disease onset, x_s and x_g indicate male (1) or female (0) and carrier (1) or non-carrier (0) of a main gene of interest, respectively.

The penetrance function for the penetrance model has the form,

$$1 - \exp\left\{-H_0(t - t_0)\exp(\beta_s x_s + \beta_g x_g)\right\},$$

where $H_0(t)$ is the cumulative baseline hazard function.

Shared frailty models

The penetrance model conditional on the frailty Z and covariates $X = (x_s, x_g)$ is assumed to have the following hazard function:

$$h(t|X, Z) = h_0(t - t_0)Z\exp(\beta_s x_s + \beta_g x_g),$$

where $h_0(t)$ is the baseline hazard function, t_0 is a minimum age of disease onset, x_s and x_g indicate male (1) or female (0) and carrier (1) or non-carrier (0) of a main gene of interest, respectively.

For example, when using a Weibull distribution for baseline hazard and a gamma distribution for frailty, the penetrance function has the form

$$1 - \left\{ 1 + \frac{\lambda^\rho (t-t_0)^\rho \exp(\beta_s x_s + \beta_g x_g)}{\kappa} \right\}^{-\kappa}.$$

Two-gene models

The penetrance curve for the two-gene model is generated by

$$1 - \exp\left\{ -H_0(t-t_0) \exp(\beta_s x_s + \beta_{g1} x_{g1} + \beta_{g2} x_{g2}) \right\},$$

where $H_0(t)$ is the cumulative baseline hazard function, x_{g1} indicates carrior (1) or non-carrior (0) of a major gene and x_{g2} indicates carrior (1) or non-carrior (0) of a second gene.

When plotting with the two-gene model, the plot will generate separate curves for mutation carriers and noncarriers, and separate curves for the second gene carriers and noncarriers.

Value

Displays plots of the penetrance functions and returns the following values:

`pen70` Penetrance estimates by age 70 specific to gender and mutation-status subgroups.

`x.age` Vector of ages of onset ranging from `agemin` to `agemax` years

`pen` Penetrance estimates computed at each age of `x.age`; if `variation = "none"` or `"frailty"`, it includes subgroups specific to gender and mutation status for major gene. If `variation = "secondgene"`, it includes subgroups specific to gender and both mutation statuses for major gene and second gene.

Author(s)

Yun-Hee Choi

See Also

`simfam`, `plot.penmodel`

Examples

```
# Penetrance function curves based on Weibull baseline hazard function

penplot(base.parms = c(0.01,3), vbeta = c(0.5, 2), variation = "none", base.dist = "Weibull",
agemin = 20, ylim = c(0,1))
```

plot.penmodel *Plot method for* penmodel

Description

Plots penetrance curves estimated from the fitted penetrance model and overlays non-parametric penetrance curves estimated from the data without proabands.

Usage

```
## S3 method for class 'penmodel'
plot(x, agemax = 80, print = TRUE, mark.time = FALSE, conf.int = FALSE,
add.KM = TRUE, MC = 100, col = c("blue", "red", "blue", "red"), lty = c(1, 1, 2, 2),
add.legend = TRUE, add.title = TRUE, xpos = "topleft", ypos = NULL,
xlab = "Age at onset", ylab = "Penetrance", ylim = NULL, main = NULL,  ...)
```

Arguments

`x` An object class of 'penmodel', a fitted model by penmodel or penmodelEM functions.

`agemax` Maximum age of disease onset or maximum age. Default is 80 years of age.

`print` Logical; if TRUE, displays parameter estimates and penetrance estimates by age 70.

`mark.time` Logical; if TRUE, curves are marked at each censoring time, otherwise, no labeling is done.

`conf.int` Logical; if TRUE, displays 95% confidence intervals for both parametric and non-parametric penetrance estimates for each subgroup and returns their lower and upper limits.

`add.KM` Logical; if TRUE, displays Kaplan-Meier curves from data.

`MC` Number of simulated samples used to calculate confidence intervals with a Monte-Carlo method. If MC = 0, no confidence intervals will be calculated. Default value is 100.

`col` Colors of lines for male carriers, female carriers, male noncarrers, and female noncarriers. Default is c("blue", "red", "blue", "red").

`lty` Types of lines for male carriers, female carriers, male noncarriers, and female noncarriers. Default is c(1, 1, 2, 2).

`add.legend` Logical; if TRUE, displays a legend in the plot.

`add.title` Logical; if TRUE, displays a title in the plot.

`xpos, ypos` Position of legend; see legend. Defaults are xpos = "topleft", ypos = NULL.

`xlab` Title for the x-axis. Default is "Age at onset".

`ylab` Title for the y-axis. Default is "Penetrance".

`ylim` Limits for the y-axis. Default is NULL. If NULL, ylim will be automatically determined.

`main` Main title of the plot. Default is `NULL`. If `NULL`, the title will be automatically created.

`...` Other parameters to be passed through to plotting functions.

Details

The 95% confidence intervals for the parametric penetrance curves are obtained based on simulations of the parameters, assuming a multivariate normal distribution for the estimated parameters with their variance-covariance matrix. See `penetrance` for more details.

Value

Returns the following summary values:

`coefficients` Parameter estimates of transformed baseline parameters (λ, ρ) and regression coefficients for gender and mutation status (β_s, β_g).

`pen70` Penetrance estimates by age 70, specific to gender and mutation-status subgroups.

`x.age` Vector of ages of onsest ranging from `agemin` to `agemax` years

`pen` Penetrance estimates at each age in `x.age`, specific to gender and mutation-status subgroups.

`lower` Lower limits of 95% confidence interval estimates for penetrance at each age in `x.age`, specific to gender and mutation status subgroups.

`upper` Upper limits of 95% confidence interval estimates for penetrance at each age in `x.age`, specific to gender and mutation status subgroups.

Author(s)

Yun-Hee Choi

See Also

`penmodel`, `print.penmodel`, `penmodelEM`, `summary.penmodel`,`print.summary.penmodel`, `simfam`

Examples

```
# Simulated family data

set.seed(4321)
fam <- simfam(N.fam = 300, design = "pop+", base.dist = "Weibull", variation = "none",
       base.parms = c(0.01,3), vbeta = c(-1.13, 2.35), allelefreq = 0.02, agemin = 20)

# Fit family data

fit <- penmodel(Surv(time, status) ~ gender + mgene, cluster = "famID", design = "pop+",
parms = c(0.01, 3, -1.13, 2.35), data = fam, base.dist = "Weibull", robust = TRUE)
```

```
# Plot penetrance function curves with 95% CIs

plot(fit, agemax = 80, conf.int = TRUE)
```

plot.simfam *Plot method for* simfam *or Plot pedigrees*

Description

Provides pedigree plots for specified families generated from simfam function with option to save plots into a pdf file.

Usage

```
## S3 method for class 'simfam'
plot(x, famid, pdf = FALSE, file = NULL, ...)
```

Arguments

x	An object of class 'simfam' created by simfam function or a data frame that has class attributes c("simfam", "data.frame").
famid	List of family IDs to plot. Default is the first family in given data set.
pdf	Logical; if TRUE, pedigree plots are saved in a pdf file. If FALSE, plot pedigrees on current plotting device. Default is FALSE.
file	File name to save the pedigree plots; Default file name is "pedigreeplot.pdf".
...	Additional arguments passed on to the plot function.

Details

Argument x can be a data frame that contains famID, indID, fatherID, motherID, gender (1 for male, 0 for female), status (1 for affected, 0 for non-affected), mgene (1 for mutation carrier, 0 for non-carrier, NA for missing), and proband (1 for proband, 0 for non-proband) and should have class attributes class(x) <- c("simfam", "data.frame").

Optionally, the data frame can contain a column named carrp.geno or carrp.pheno to replace missing values in mgene with their carrier probabilities.

Value

Returns pedigree plots for specified families created by simfam function or for the data frame provided along with the affection and carrier mutation statuses of family members. Probands from each pedigree are indicated using red color.

When object inlcudes carrp.geno and/or carrp.pheno generated by carrierprob function, the plot function displays the carrier probabilities for those with missing carrier status.

See Also

simfam, summary.simfam, carrierprob

Examples

```
# Simulated family data

set.seed(4321)
fam <- simfam(N.fam = 200, design = "pop+", base.dist = "Weibull", allelefreq = 0.02,
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35), agemin = 20)

# Pedigree plots for first three simulated families

plot(fam, famid = c(1:3))
```

print.penmodel *Print method for* penmodel.

Description

Prints a summary of parameter estimates of a fitted penetrance model.

Usage

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

Arguments

x	An object class of 'penmodel', a fitted model by penmodel or penmodelEM functions.
digits	Number of significant digits to use when printing.
...	Further arguments passed to or from other methods.

Value

Prints a short summary of the model and model fit.

Returns an object of class 'penmodel'.

Author(s)

Yun-Hee Choi

See Also

penmodel, penmodelEM, summary.penmodel, print.summary.penmodel, plot.penmodel

print.summary.penmodel

Print method for `summary.penmodel` *of a fitted penetrance model.*

Description

Prints a short summary of parameter and penetrance estimates of a `'summary.penmodel'` object.

Usage

```
## S3 method for class 'summary.penmodel'
print(x, digits = max(3, getOption(”digits”) - 3),
  signif.stars=TRUE, ...)
```

Arguments

`x` An object class of `'summary.penmodel'`, a result of a call to `summary.penmodel`.

`digits` Number of significant digits to use when printing.

`signif.stars` Logical; if `TRUE`, provides stars to hightlight significant p-values. Default is `TRUE`.

`...` Further arguments passed to or from other methods.

Value

Prints a summary of parameter estimates, their standard errors, t-statistics and corresponding two-sided p-values and additionally indicates significance stars if `signif.stars` is `TRUE`.

Also prints penetrance estimates by age 70 specific to gender and mutation-status subgroups along with their standard errors and 95% confidence intervals.

Returns an object of class `'summary.penmodel'`.

Author(s)

Yun-Hee Choi

See Also

`penmodel`, `penmodelEM`, `print.penmodel`, `summary.penmodel`

simfam *Generate familial time-to-event data*

Description

Generates familial time-to-event data for specified study design, genetic model and source of residual familial correlation; the generated data frame also contains family structure (individual's id, father id, mother id, relationship to proband, generation), gender, current age, genotypes of major or second genes.

Usage

```
simfam(N.fam, design = "pop", variation = "none", interaction = FALSE, depend = NULL,
       base.dist = "Weibull", frailty.dist = NULL, base.parms, vbeta,
       allelefreq = c(0.02, 0.2), dominant.m = TRUE, dominant.s = TRUE,
       mrate = 0, hr = 0, probandage = c(45, 2), agemin = 20, agemax = 100)
```

Arguments

`N.fam` Number of families to generate.

`design` Family based study design used in the simulations. Possible choices are: `"pop"`, `"pop+"`, `"cli"`, `"cli+"` or `"twostage"`, where `"pop"` is for the population-based design that families are ascertained by affected probands, `"pop+"` is similar to `"pop"` but with mutation carrier probands, `"cli"` is for the clinic-based design that includes affected probands with at least one parent and one sib affected, `"cli+"` is similar to `"cli"` but with mutation carrier probands and `"twostage"` for two-stage design that randomly samples families from the population in the first stage and oversamples high risk families in the second stage that include at least two affected members in the family. Default is `"pop"`.

`variation` Source of residual familial correlation. Possible choices are: `"frailty"` for frailty shared within families, `"secondgene"` for second gene variation, or `"none"` for no residual familial correlation. Default is `"none"`.

`interaction` Logical; if `TRUE`, allows the interaction between gender and mutation status. Default is `FALSE`.

`depend` Variance of the frailty distribution. Dependence within families increases with depend value. Default is `NULL`. Value should be specified as a positive real number when `variation="frailty"`.

`base.dist` Choice of baseline hazard distribution. Possible choices are: `"Weibull"`, `"loglogistic"`, `"Gompertz"`, `"lognormal"` `"gamma"`, `"logBurr"`. Default is `"Weibull"`.

`frailty.dist` Choice of frailty distribution. Possible choices are: `"gamma"` or `"lognormal"` when `variation="frailty"`. Default is `NULL`.

`base.parms` Vector of parameter values for the specified baseline hazard function. `base.parms=c(lambda, rho)` should be specified for `base.dist="Weibull"`, `"loglogistic"`, `"Gompertz"`, `"gamma"`, and `"lognormal"`. For `base.dist="logBurr"`, three parameters should be specified `base.parms = c(lambda, rho, eta)`.

`vbeta` Vector of regression coefficients for gender, majorgene, interaction between gender and majorgene (if `interaction = TRUE`), and secondgene (if `variation = "secondgene"`).

`allelefreq` Vector of population allele frequencies of major and second disease gene alleles. Frequencies must be between 0 and 1. Default frequencies are 0.02 for major gene allele and 0.2 for second gene allele, `allelefreq = c(0.02, 0.2)`.

`dominant.m` Logical; if `TRUE`, the genetic model of major gene is dominant, otherwise recessive.

`dominant.s` Logical; if `TRUE`, the genetic model of second gene is dominant, otherwise recessive.

`mrate` Proportion of missing genotypes, value between 0 and 1. Default value is 0.

`hr` Proportion of high risk families, which include at least two affected members, to be sampled from the two stage sampling. This value should be specified when `design="twostage"`. Default value is 0. Value should lie between 0 and 1.

`probandage` Vector of mean and standard deviation for the proband age. Default values are mean of 45 years and standard deviation of 2 years, `probandage = c(45, 2)`.

`agemin` Minimum age of disease onset or minimum age. Default is 20 years of age.

`agemax` Maximum age of disease onset or maximum age. Default is 100 years of age.

Details

The `design` argument defines the type of family based design to be simulated. Two variants of the population-based and clinic-based design can be chosen: `"pop"` when proband is affected, `"pop+"` when proband is affected mutation carrier, `"cli"` when proband is affected and at least one parent and one sibling are affected, `"cli+"` when proband is affected mutation-carrier and at least one parent and one sibling are affected. The two-stage design, `"twostage"`, is used to oversample high risk families, where the proportion of high risks families to include in the sample is specified by `hr`. High risk families often include multiple (at least two) affected members in the family.

The ages at onset are generated from the following penetrance models depending on the choice of `variation = "none", "frailty", "secondgene"`.. When `variation = "none"`, the ages at onset are independently generated from the proportional hazard model conditional on the gender and carrier status of major gene mutation, $X = (x_s, x_g)$.

The ages at onset correlated within families are generated from the shared frailty model (codevariation = "frailty") or the two-gene model (codevariation = "secondene"), where the residual familial correlation is induced by a frailty or a second gene, respectively, shared within the family.

The proportional hazard model

$$h(t|X) = h_0(t - t_0)\exp(\beta_s x_s + \beta_g x_g),$$

where $h_0(t)$ is the baseline hazard function, t_0 is a minimum age of disease onset, x_s and x_g indicate male (1) or female (0) and carrier (1) or non-carrier (0) of a main gene of interest, respectively.

The shared frailty model

$$h(t|X, Z) = h_0(t - t_0) Z \exp(\beta_s x_s + \beta_g x_g),$$

where $h_0(t)$ is the baseline hazard function, t_0 is a minimum age of disease onset, Z represents a frailty shared within families and follows either a gamma or log-normal distribution, x_s and

x_g indicate male (1) or female (0) and carrier (1) or non-carrier (0) of a main gene of interest, respectively.

The two-gene model

$$h(t|X, Z) = h_0(t - t_0)Z \exp(\beta_s x_s + \beta_1 x_1 + \beta_2 x_2),$$

where x_1, x_2 indicate carriers (1) and non-carriers (0) of a major gene and of second gene mutation, respectively.

The current ages for each generation are simulated assuming normal distributions. However, the probands' ages are generated using a left truncated normal distribution as their ages cannot be less than the minimum age of onset. The average age difference between each generation and their parents is specified as 20 years apart.

Note that simulating family data under the clinic-based designs (`"cli"` or `"cli+"`) or the two-stage design can be slower since the ascertainment criteria for the high risk families are difficult to meet in such settings. Especially, `"cli"` design could be slower than `"cli+"` design since the proband's mutation status is randomly selected from a disease population in `"cli"` design, so his/her family members are less likely to be mutation carriers and have less chance to be affected, whereas the probands are all mutation carriers, their family members have higher chance to be carriers and affected by disease. Therefore, `"cli"` design requires more iterations to sample high risk families than `"cli+"` design. All designs simulations that include `variation = "frailty"` could be also slower in order to generate families with specific familial correlations induced by the chosen frailty distribution.

Value

Returns an object of class `'simfam'`, a data frame which contains:

`famID`	Family identification (ID) numbers.
`indID`	Individual ID numbers.
`gender`	Gender indicators: 1 for males, 0 for females.
`motherID`	Mother ID numbers.
`fatherID`	Father ID numbers.
`proband`	Proband indicators: 1 if the individual is the proband, 0 otherwise.
`generation`	Individuals generation: 1=parents of probands,2=probands and siblings, 3=children of probands and siblings.
`majorgene`	Genotypes of major gene: 1=AA, 2=Aa, 3=aa where A is disease gene.
`secondgene`	Genotypes of second gene: 1=BB, 2=Bb, 3=bb where B is disease gene.
`ageonset`	Ages at disease onset in years.
`currentage`	Current ages in years.
`time`	Ages at disease onset for the affected or ages of last follow-up for the unaffected.
`status`	Disease statuses: 1 for affected, 0 for unaffected (censored).
`mgene`	Major gene mutation indicators: 1 for mutated gene carriers, 0 for mutated gene noncarriers, or `NA` if missing.
`relation`	Family members' relationship with the proband:

1	Proband (self)
2	Brother or sister
3	Son or daughter
4	Parent
5	Nephew or niece
6	Spouse
7	Brother or sister in law

`fsize` Family size including parents, siblings and children of the proband and the siblings.

`naff` Number of affected members in family.

`weight` Sampling weights.

Author(s)

Yun-Hee Choi, Wenqing He

References

Choi, Y.-H., Kopciuk, K. and Briollais, L. (2008) Estimating Disease Risk Associated Mutated Genes in Family-Based Designs, Human Heredity 66, 238-251

Choi, Y.-H. and Briollais (2011) An EM Composite Likelihood Approach for Multistage Sampling of Family Data with Missing Genetic Covariates, Statistica Sinica 21, 231-253

See Also

summary.simfam, plot.simfam, penplot

Examples

```
## Example 1: simulate family data from population-based design using
#  a Weibull distribution for the baseline hazard and inducing
#  residual familial correlation through a shared gamma frailty.

set.seed(4321)
fam <- simfam(N.fam = 10, design = "pop+", variation = "frailty",
       base.dist = "Weibull", frailty.dist = "gamma", depend=1,
       allelefreq = 0.02, base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35))

head(fam)

## Not run:
  famID indID gender motherID fatherID proband generation majorgene secondgene
1     1     1      1        0        0       0          1         2          0
2     1     2      0        0        0       0          1         2          0
3     1     3      0        2        1       1          2         2          0
4     1     4      1        0        0       0          0         3          0
5     1     9      0        3        4       0          3         2          0
6     1    10      1        3        4       0          3         3          0
```

```
   ageonset currentage     time status mgene relation fsize naff weight
1 103.76925   69.19250 69.19250      0     1        4    18    2      1
2  64.88982   67.31119 64.88982      1     1        4    18    2      1
3  45.84891   47.57119 45.84891      1     1        1    18    2      1
4 269.71990   47.37403 47.37403      0     0        6    18    2      1
5  69.78355   27.80081 27.80081      0     1        3    18    2      1
6 192.09392   25.34148 25.34148      0     0        3    18    2      1

## End(Not run)

summary(fam)

plot(fam, famid = c(1:2)) # pedigree plots for families with IDs = 1 and 2

## Example 2: simulate family data from two stage design to include
#  30% of high risk families in the sample.

set.seed(4321)
fam <- simfam(N.fam = 50, design = "twostage", variation = "none", base.dist = "Weibull",
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35), hr = 0.3, allelefreq = 0.02)

summary(fam)
```

`summary.penmodel` *Summary method for class* `penmodel`

Description

Provides a summary of a fitted penetrance model.

Usage

```
## S3 method for class 'penmodel'
summary(object, correlation=FALSE, ...)
```

Arguments

`object` An object class of `'penmodel'`, a fitted model by `penmodel` or `penmodelEM` functions.

`correlation` Logical; if `TRUE`, returns the correlation matrix of the estimated parameters. Default is `FALSE`.

`...` Further arguments passed to or from other methods.

Value

Returns the object of class `'summary.penmodel'`, including the following summary values:

`estimates` List of parameter estimates of transformed baseline parameters and regression coefficients, their standard errors, their robust standard errors if `robust=TRUE` was selected when fitting the penetrance model, t-statistics and corresponding two-sided p-values.

`varcov` Variance-covariance matrix of the parameter estimates.

`varcov.robust` Robust variance-covariance matrix of the parameter estimates if `robust = TRUE` was selected when fitting the penetrance model.

`correlation` Correlation matrix obtained from the variance-covariance matrix.

`correlation.robust`
Correlation matrix obtained from the robust variance-covariance matrix if `robust = TRUE` was selected when fitting the penetrance model.

Author(s)

Yun-Hee Choi

See Also

`penmodel`, `penmodelEM`, `print.penmodel`, `print.summary.penmodel` `plot.penmodel`

Examples

```
# Simulated family data

set.seed(4321)
fam <- simfam(N.fam = 200, design = "pop+", variation = "none", base.dist = "Weibull",
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35), agemin = 20, allelefreq = 0.02)

# Penetrance model fit for the simulated family data

fit <- penmodel(Surv(time, status) ~ gender + mgene, cluster = "famID",
      parms=c(0.01, 3, -1.13, 2.35), data = fam, design = "pop+", base.dist = "Weibull")

# Summary of the model parameter and penetrance estimates from model fit

summary(fit)

## Not run:
Estimates:
            Estimate Std. Error t value Pr(>|t|)
log(lambda)   -4.531    0.08583 -52.793  0.01206 *
log(rho)       1.113    0.04688  23.737  0.02680 *
gender        -1.302    0.19233  -6.768  0.09339 .
mgene          2.349    0.23825   9.859  0.06436 .
Signif. codes:   0 '***'  0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## End(Not run)
```

summary.simfam *Summary method for* simfam

Description

Provides a summary of simulated data.

Usage

```
## S3 method for class 'simfam'
summary(object, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

object	An object class of 'simfam' generated from simfam function
digits	Number of significant digits to use when printing.
...	Further arguments passed to or from other methods.

Value

Displays a summary of simulated data and returns the following values:

num.fam	Number of families simulated.
avg.num.affected	Average number of affected individuals per family.
avg.num.carriers	Average number of mutation carriers per family.
avg.family.size	Average family size.
ave.ageonset	Average age of onset for affected individuals.

Author(s)

Yun-Hee Choi

See Also

simfam

Examples

```
set.seed(4321)
fam <- simfam(N.fam = 50, design = "pop", variation = "none", base.dist = "Weibull",
       base.parms = c(0.01, 3), vbeta = c(-1.13, 2.35))

summary(fam)
## Not run:
```

```
Study design:                           pop
Baseline distribution:                  Weibull
Number of families:                     50
Average number of affected per family:  1.24
Average number of carriers per family:  1.3
Average family size:                    17.02
Average age of onset for affected:      40.08
Sampling weights used:                  1

## End(Not run)
```

Index