Package 'FamEvent'

March 26, 2017

Type Package

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

Version 1.3

Date 2017-03-25

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

Depends R(>= 3.0.0), base, MASS, truncnorm, survival, kinship2

Repository CRAN

Date/Publication 2017-03-26 07:30:45 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

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(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam,
       design="pop+", base.dist="Weibull")

# 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 or "model" using the parametric model fit; see details. 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 status 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="model" 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.

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(parms=c(0.01, 3), vbeta=c(-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)
```

---

penci *Estimate the confidence intervals for the penetrances*

Description

Estimates the simulation-based confidence intervals for the penetrances and the standard errors of the penetrance estimates at a given age, specific to gender and mutation status subgroups.

Usage

```
penci(est, cov, age=70, base.dist="Weibull", frailty.dist=NULL, agemin, n=1000)
```

Arguments

est	Parameter estimates of transformed baseline parameters (λ, ρ) and regression coefficients for gender and mutation status (β_s, β_g) for the assumed penetrance model.
cov	Variance-covariance matrix of the parameter estimates.
age	Specified age at which the penetrance is computed.

base.dist	Choice of baseline hazard distribution for the penetrance function. Possible choices are "Weibull", "loglogistic", "Gompertz", "lognormal", or "gamma". Default is "Weibull".
frailty.dist	Choice of frailty distribution, either "gamma" or "lognormal", if the penetrance function is based on a shared frailty model. Otherwise, frailty.dist = NULL is set as default.
agemin	Minimum age that the pentrance function starts.
n	Number of Monte-Carlo simulations for calculating standard errors and 95% confidence intervals for the penetrance estimate at a given age. Default value is n = 1000.

Details

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 n sets of the parameters are generated from the multivariate normal distribution with the parameter estimates and their variance and 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 an object including the following values:

Estimate	Penetrance estimates (%) at the specified age, specific to gender and mutation status subgroups, based on the assumed penetrance model.
SE	Simulation-based standard errors of the penetrance estimates, specific to gender and mutation status subgroups.
lower	Simulation-based 2.5th percentile of the penetrance estimates, specific to gender and mutation status subgroups.
upper	Simulation-based 97.5th percentile of the penetrance estimates, specific to gender and mutation status subgroups.

Author(s)

Yun-Hee Choi

See Also

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

Examples

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

fam <- simfam(N.fam=300, 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(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam,
                design="pop+", base.dist="Weibull")

# 95% confidence intervals for the penetrance at age 50 based on 1000 simulations

penci(fit$coefficients, fit$varcov, age=50, base.dist="Weibull", agemin=20, n=1000)
```

---

penf *Penetrance function*

Description

Calculates the penetrance for a given age, gender and mutation status based on the assumed penetrance model.

Usage

```
penf(est, age, sex, mut, base.dist, frailty.dist=NULL, agemin)
```

Arguments

`est`	Parameter estimates for the assumed penetrance model including the transformed baseline parameters (λ, ρ), regression coefficients for gender and mutation status (β_s, β_g) and a frailty parameter (κ) if `frailty.dist` is specified.
`age`	Vector of ages or a single value of age at which the penetrance function is evaluated.
`sex`	1 for male, 0 for female.
`mut`	1 for mutation carrier, 0 for mutation noncarrier.
`base.dist`	Choice of baseline hazard distribution for the penetrance function. Possible choices are `"Weibull"`, `"loglogistic"`, `"Gompertz"`, `"lognormal"`, or `"gamma"`. Default is `"Weibull"`.
`frailty.dist`	Choice of frailty distribution, either `"gamma"` or `"lognormal"`, if the penetrance function is based on a shared frailty model. Otherwise, `frailty.dist = NULL` is set as default.
`agemin`	Minimum age that the pentrance function starts.

Details

The penetrance function is defined as the probability of developing disease by time t given gender (x_s) and mutation status (x_g),

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

where $S(t; x_s, x_g)$ is the survival distribution based on a proportional hazards model with a specified baseline hazard distribution or based on a shared frailty mode with specified frailty and baseline hazard distributions.

Proportional hazards model:

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

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).

Shared frailty model:

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

where z is the shared frailty whose distribution is specified by `frailty.dist` with associated parameter κ. The marginal survival function is obtained by integrating the conditional survival distribution, $S(s; z, x_s, x_g)$, over frailty distribution, i.e.,

$$S(t; x_s, x_g) = \int_0^\infty S(s; z, x_s, x_g) dG(s)$$

where $G(s)$ is the cumulative distribution function for the frailty.

Value

Returns a penetrance value evaluated at given age, gender and mutation status based on the specified penetrance model.

Author(s)

Yun-Hee Choi

See Also

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

Examples

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

fam <- simfam(N.fam=300, 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(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam,
                 design="pop+", base.dist="Weibull")

# Computing the penetrance at age 50 for male (sex=1) carriers (mut=1) from the assumed
# penetrance model based on Weibull baseline hazard.

penf(fit$coefficients, age=50, sex=1, mut=1, base.dist="Weibull", agemin=20)
```

`penmodel` *Estimate the penetrance model and penetrance curves*

Description

Fits a penetrance model for family data based on a prospective likelihood with ascertainment correction and provides parameter estimates as well as the gender- and mutation-specific penetrance estimates.

Usage

```
penmodel(parms, vbeta, data, design="pop", base.dist="Weibull", robust=FALSE)
```

Arguments

`parms`	Vector of initial values for baseline parameters. `parms=c(lambda, rho)`, where lambda and rho are the initial values for the scale and shape parameters, respectively. For the `"lognormal"` baseline distribution, `rho` > 0; for the other baselinse distributions, `lambda` > 0 and `rho` > 0.
`vbeta`	Vector of initial values for regression coefficients for gender and majorgene; `vbeta=c(beta.s, beta.g)`.
`data`	Data frame generated from `simfam` or data frame containing specific variables: `famID`, `indID`, `generation`, `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 distribution to fit. Possible choices are: "Weibull", "loglogistic", "Gompertz", "lognormal", or "gamma". Default is "Weibull".
robust	Logical; if TRUE, use robust 'sandwich' standard errors and variance covariance matrix, otherwise use conventional standard errors and variance covariance matrix.

Details

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

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

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).

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 from 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 based on the composite likelihood approach (Choi and Briollais, 2011)

Transformed baseline parameters (λ, ρ) were used for estimation; log tranformation was applied to both scale and shape parameters for "Weibull", "loglogistic", "Gompertz" and "gamma" baseline distributions. For "lognormal" baseline distribution, the log transformation was applied only to shape parameter ρ, not to λ which represents the location parameter in log-normal distribution.

Calculations of standard errors and 95% confidence intervals for penetrance estimates by age 70 were based on the penetrances obtained from 1000 Monte-Carlo simulations of the estimated penetrance model; for more details, see penci.

Value

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

coefficients	Parameter estimates of transformed baseline parameters (λ, ρ) and regression coefficients for gender and mutation status (β_s, β_g).
varcov	Variance covariance matrix of parameter estimates. If robust=TRUE, robust 'sandwich' variance covariance matrix is returned.
se	Standard errors of parameter estimates. If robust=TRUE, robust 'sandwich' standard errors are returned.
pen70.est	Penetrance estimates by age 70 specific to gender and mutation-status subgroups.
pen70.se	Standard errors of penetrance estimates by age 70 specific to gender and mutation-status subgroups.
pen70.ci	95% confidence interval for penetrance estimates by age 70 specific to gender and mutation-status subgroups.
ageonset	Vector of ages of onset ranging from agemin to 90 years.
pen.maleCarr	Vector of penetrance estimates for male carriers from agemin to 90 years.
pen.femaleCarr	Vector of penetrance estimates for female carriers from agemin to 90 years.

`pen.maleNonCarr`
: Vector of penetrance estimates for male non-carriers from `agemin` to 90 years.

`pen.femaleNonCarr`
: Vector of penetrance estimates for female non-carriers from `agemin` to 90 years.

`logLik`
: Loglikelihood value for the fitted penetrance 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`, `penci`, `penf`

Examples

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

fam <- simfam(N.fam=300, 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(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam,
       design="pop+", base.dist="Weibull")

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

summary(fit)

# Generate the lifetime penetrance curves from model fit for specific gender and
# mutation status groups along with their non-parametric penetrance curves

plot(fit)
```

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 and corresponding gender- and genotype-specific penetrance estimates.

Usage

```
penmodelEM(parms, vbeta, data, design="pop", base.dist="Weibull",
           robust=FALSE, method="data", mode="dominant", q=0.02)
```

Arguments

parms	Vector of initial values for baseline parameters. parms=c(lambda, rho), where lambda and rho are the initial values for the scale and shape parameters, respectively. For the "lognormal" baseline distribution, rho > 0; for the other baselinse distributions, lambda > 0 and rho > 0.
vbeta	Vector of initial values for the regression coefficients for gender and majorgene, vbeta=c(beta.s, beta.g).
data	Data frame generated from simfam or data frame containing specific variables: famID, indID, generation, 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 distribution to fit. Possible choices are: "Weibull", "loglogistic", "Gompertz", "lognormal", or "gamma". Default is "Weibull".
robust	Logical; if TRUE, use robust 'sandwich' standard errors and variance matrix, otherwise use conventional standard errors and variance matrix.
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 probabilies 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

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

where X_{fi} represents the mutation carrier status and Y_{fi} represents the phenotype (t_{fi}, δ_{fi}) in terms of age at onset t_{fi} and disease status δ_{fi} for individual i in family f and X_f^o represents the observed genotypes in family f.

Then, we obtain the conditional expectation of the log-likelihood function 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).

Transformed baseline parameters (λ, ρ) were used for estimation; see `penmodel` for details.

Value

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

`coefficients` Parameter estimates of transformed baseline parameters (λ, ρ) and regression coefficients for gender and mutation status (β_s, β_g) including their standard errors and also their robust standard errors.

`varcov` Variance covariance matrix of parameter estimates. If `robust=TRUE`, robust 'sandwich' variance covariance matrix is returned.

`se` Standard errors of parameter estimates. If `robust=TRUE`, robust 'sandwich' standard errors are returned.

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

pen70.se	Standard errors of penetrance estimates by age 70 specific to gender and mutation-status subgroups.
pen70.ci	95% confidence interval estimates of penetrance by age 70 specific to gender and mutation-status subgroups.
ageonset	Vector of ages of onset ranging from agemin to 90 years.
pen.maleCarr	Vector of penetrance estimates for male carriers from agemin to 90 years.
pen.femaleCarr	Vector of penetrance estimates for female carriers from agemin to 90 years.
pen.maleNoncarr	Vector of penetrance estimates for male non-carriers from agemin to 90 years.
pen.femaleNoncarr	Vector of penetrance estimates for female non-carriers from agemin to 90 years.
logLik	Loglikelihood value for the fitted penetrance 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 30% of members missing their genetic information.

fam <- simfam(N.fam=100, design="pop+", base.dist="Weibull", base.parms=c(0.01,3),
       vbeta=c(-1.13, 2.35), agemin=20, allelefreq=0.02, mrate=0.3)

# EM algorithm for fitting family data with missing genotypes

fit <- penmodelEM(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam, design="pop+",
       base.dist="Weibull", method="mendelian", mode="dominant", q=NULL)

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

summary(fit)
```

```
# Generate the lifetime penetrance curves from model fit for gender and
# mutation status groups along with their non-parametric penetrance curves
# based on observed data

plot(fit)
```

---

penplot *Plot penetrance functions*

Description

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

Usage

```
penplot(base.parms, vbeta, variation="none", base.dist="Weibull",
       frailty.dist=NULL, depend=1, agemin=20, print=TRUE, ...)
```

Arguments

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.
vbeta	Vector of regression coefficients for gender and majorgene vbeta=c(beta.s, beta.g). If variation="secondgene", specify regression coefficient for second gene in vbeta=c(beta.s, beta.g1, beta.g2).
base.dist	Choice of baseline hazard distribution. Possible choices are: "Weibull", "loglogistic", "Gompertz", "lognormal", or "gamma". 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.
variation	Source of residual familial correlation. Possible choices are "frailty" for frailty shared within families, "secondgene" for second gene shared within families, or "none" for no residual familial correlation. Default is "none"
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.
print	Logical; if TRUE, prints the penetrance values by age 70 from the assumed model.
...	Other parameters to be passed through to plotting functions.

Details

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

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

The penetrance curve for the second gene model is generated by

$$1 - \exp\left\{-\lambda^\rho (t - t_0)^\rho \exp(\beta_g x_g + \beta_{g1} x_{g1} + \beta_{g2} x_{g2})\right\}$$

where 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 second gene model, the plot will generate separate curves for mutation carriers and noncarriers, and seperate 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`	Vetor of ages of onsest ranging from `agemin` to 80 years
`pen`	Lists of penetrance estimates computed at each age of `x.age`; if `variation = "none"` or `"frailty"`, lists include subgroups specific to gender and mutation status for major gene. If `variation = "secondgene"`, lists include 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), base.dist="Weibull", agemin=20)
```

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, print=TRUE, mark.time=FALSE, conf.int=FALSE, ...)
```

Arguments

x	An object class of penmodel, a fitted model by penmodel or penmodelEM functions.
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.
...	Other parameters to be passed through to plotting functions.

Details

The 95% confidence intervals for the parametric penetrance curves were obtained based on 1000 simulations of the parameters, assuming a multivariate normal distribution for the estimated parameters with their variance covariance matrix. See penci for more details.

Value

It displays 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	Vetor of ages of onsest ranging from agemin to 80 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, penci, penf

Examples

```
# Simulated family data

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(parms=c(0.01, 3), vbeta=c(-1.13, 2.35), data=fam,
       design="pop+", base.dist="Weibull")

# Plot penetrance function curves

plot(fit)
```

---

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 or for dataframe containing required .

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 dataframe 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

fam <- simfam(N.fam=300, 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), pdf=TRUE, file="pedigrees.pdf")
```

---

print.penmodel *Print method for* penmodel.

Description

Prints a summary of parameter and penetrance 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	the 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.
...	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", depend=1,
       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, age1=c(65,2.5), age2=c(45,2.5), agemin=20)
```

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".
depend	Variance of the frailty distribution. Dependence within families increases with depend value. Default value is 1.
base.dist	Choice of baseline hazard distribution. Possible choices are: "Weibull", "loglogistic", "Gompertz", "lognormal", or "gamma". 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.
`vbeta`	Vector of parameter values for gender, majorgene, and 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.
`age1`	Vector of mean and standard deviation for the current age of generation 1 or grandparents. Default values are mean of 65 years and standard deviation of 2.5 years, `age1=c(65,2.5)`.
`age2`	Vector of mean and standard deviation for the current age of generation 2 or proband generation. Default values are mean of 45 years and standard deviation of 2.5 years, `age2=c(45,2.5)`.
`agemin`	Minimum age of disease onset. Default is 20 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.

Age at onset is generated from the penetrance model where residual familial correlation is induced by either a latent random variable called "frailty"" or a second gene shared by family members.

The penetrance model with a shared frailty model has the form

$$h(t|Z) = h_0(t - t_0)Z \exp(\beta_s x_s + \beta_{g1} x_{g1})$$

where Z represents a frailty shared within families and follows either a gamma or log-normal distribution; t_0 is a minimum age of disease onset; x_s indicates males (1) and females (0) and x_{g1} indicates carriers (1) and non-carriers (0) of major gene mutation.

The penetrance model with a second gene variation has the form

$$h(t|Z) = h_0(t - t_0)\exp(\beta_s x_s + \beta_{g1} x_{g1} + \beta_{g2} x_{g2})$$

where x_{g2} indicates carriers (1) and non-carriers (0) of a second gene mutation.

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 mean age difference between each generation and their parents is specified as at least 20 years apart.

Value

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

`famID`	Family identification number (id).
`indID`	Individual id.
`gender`	Gender indicator: 1 for males, 0 for females.
`motherID`	Mother id number.
`fatherID`	Father id number.
`proband`	Proband indicator: 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`	Genotype of major gene: 1=AA, 2=Aa, 3=aa where A is disease gene.
`secondgene`	Genotype of second gene: 1=BB, 2=Bb, 3=bb where B is disease gene.
`ageonset`	Age at disease onset.
`currentage`	Current age.
`time`	Minimum time between current age and age at onset.
`status`	Disease status: 1 for affected and 0 for unaffected (censored).
`mgene`	Carrier status of major gene which can possibly be missing: 1 for carrier, 2 for non-carrier, NA for missing carrier status.
`relation`	Family members' relationship with the proband is as follows:

1	Proband (self)
2	Brother or sister
3	Son or daughter
4	Parent
5	Nephew or niece
6	Husband
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.

fam <- simfam(N.fam=100, 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         3          0
 3     1     3      0        2        1       1          2         2          0
 4     1     4      1        0        0       0          0         3          0
 5     1     7      0        3        4       0          3         2          0
 6     1     8      1        3        4       0          3         3          0
   ageonset currentage time status mgene relation fsize naff weight
 1       70         68   68      0     1        4    11    1      1
 2      110         68   68      0     0        4    11    1      1
 3       36         40   36      1     1        1    11    1      1
 4      212         50   50      0     0        6    11    1      1
 5       79         19   19      0     1        3    11    1      1
 6      169         16   16      0     0        3    11    1      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.
```

```
fam <- simfam(N.fam=100, design="twostage", variation="frailty",
       base.dist="Weibull", frailty.dist="gamma", depend=1, hr=0.3,
       base.parms=c(0.01,3), vbeta=c(-1.13, 2.35), 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.
`...`	Further arguments passed to or from other methods.

Value

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

`coefficients`	4 x 4 matrix with columns for parameter estimates of transformed baseline parameters (λ, ρ) and regression coefficients for gender and mutation status (β_s, β_g), their standard errors (or robust standard errors if `robust=TRUE` was selected when fitting the penetrance model), t-statistics and corresponding two-sided p-values.
`varcov`	4 x 4 variance covariance matrix of the parameter estimates.
`correlation`	Correlation matrix corresponding to the specified variance covariance matrix, if `correlation=TRUE` is specified.
`pen70`	Penetrance estimates by age 70 specific to gender and mutation-status subgroups including their standard errors and 95% confidence intervals.

Author(s)

Yun-Hee Choi

See Also

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

Function coef will extract the matrix of coefficients with standard errors, t-statistics and p-values.

Examples

```
# Simulated family data

fam <- simfam(N.fam=300, 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(parms=c(0.01, 3), vbeta=c(-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:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
log(lambda)   -4.637    0.07989 -58.048  0.01097 *
log(rho)       1.094    0.03910  27.980  0.02274 *
beta.sex      -1.315    0.16298  -8.066  0.07852 .
beta.gene      2.545    0.21423  11.881  0.05346 .
Signif. codes:   0 '***'  0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Penetrance (%) by age 70:
         Male Carrier Female Carrier Male Noncarrier Female Noncarrier
Estimate        32.47          76.82            3.03             10.83
SE               3.85           3.22            0.74              2.17

95% Confidence intervals on the penetrances:
           Male Carrier Female Carrier Male Noncarrier Female Noncarrier
lowerlimit        25.59          70.43            1.94              7.43
upperlimit        40.70          83.08            4.85             15.87

## 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

```
fam <- simfam(N.fam=100, design="pop", variation="frailty", depend=1,
       frailty.dist="gamma", 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
Frailty distribution:                  gamma
Number of families:                    100
Average number of affected per family:  1.28
Average number of carriers per family:  2.3
Average family size:                   16.02
Average age of onset for affected:     39.73
Sampling weights used:                 1
```

```
## End(Not run)
```

Index