Package 'stacd'

October 16, 2012

Type Package

Title The STACD model for two-locus genetic interactions

Version 0.1-2

Date 2012-10-16

Author Raymond Walters, Charles Laurin, Gitta Lubke

Maintainer Raymond Walters <rwalters@nd.edu>

Description

The 'stacd' package provides a set of utilites for using the STACD model. The STACD model is designed to model interactions between two biallelic loci using an intuitive parameterization with interpretable coefficients. The model is applicable to both dichotomous and contiuous phenotypes. Functions are included for variance decomposition, data generation, power analysis, and coversion between the STACD model and other common interaction models.

License GPL (>=2)

LazyLoad yes

R topics documented:

`logodds.to.stacd` *Convert penetrance or log odds model to STACD parameters*

Description

Two-locus interaction models are commonly presented as using penetrances or log odds ratios. These function expresses any given two-locus penetrance or log odds ratio model in terms of the STACD model parameters. Additional information about the model is then provided based on the STACD model parameters.

Usage

```
pen.to.stacd(pen.M, pa = NULL, pb = NULL)
logodds.to.stacd(logodds.M, pa = NULL, pb = NULL)
```

Arguments

`pen.M`	A 3x3 matrix containing the penetrances (*P(y=1)* conditional on the genotypes) for the model. The columns of the matrix should correspond to the genotype at the first locus, and the rows should correspond to the genotype at the second locus.
`logodds.M`	A 3x3 matrix containing the log odds ratios (*ln[P(y=1)/(1-P(y=1))]* conditional on the genotypes) for the model. The columns of the matrix should correspond to the genotype at the first locus, and the rows should correspond to the genotype at the second locus.
`pa`	(optional) allele frequency for the first locus. If omitted, defaults to .5 with a warning.
`pb`	(optional) allele frequency for the second locus. If omitted, defaults to .5 with a warning.

Details

STACD model parameters are computed using just the provided penetrance model. All other information about the model (marginals, variance decomposition) requires that the allele frequencies also be specified. If allele frequencies `pa` and/or `pb` are not given, then an allele frequncy of 0.5 is assumed, with a warning.

The two loci for the penetrance model are assumed to be biallelic, unlinked, and under Hardy-Weinberg equilibrium. The variance decomposition assumes the dichotomous phenotype results from thresholding a continuous latent variable with logistically-distributed errors.

Value

`Betas`	STACD model parameters
`Variance_Explained`	a list with elements: `Multiple_R_Sq` - total proportion of variance explained by ther model, as per McKelvey & Zavoina (1975). `Parameter_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the STACD paramters `Cockerham_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the conventional decomposition given by Cockerham (1954).
`Penetrances`	the provided 3x3 penetrance matrix
`Marginals`	length 3 vectors of the marginal pentrance conditional on the first locus and on the second locus
`Prevalence`	expected prevalence of the dichotmoous phenotype
`Var_ystar`	variance of the continuous variable underlying the phenotype

Author(s)

Raymond Walters, Charles Laurin

References

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39: 859-882.

McKelvey, R., and Zavoina, W. (1975). A Statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103-120.

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

See Also

To compute a penetrance model from STACD parameters see `stacd.model`

Examples

```
# Example from Walters et al using log odds model
# adapted from data in Bolton et al 2010
exvarex <- logodds.to.stacd(logodds.M=matrix(c(0,0,0, 0,0,.25, 0,0,1),3,3,byrow=TRUE),
                            pa=0.3, pb=0.3)
print(exvarex$Betas)
print(exvarex$Variance_Explained)

# Example using the Modifying Effect Model
pen.M <- matrix(c(.01,.01,.01,
.01,.01,.02,
.02,.02,.02),3,3,byrow=TRUE)
pen.to.stacd(pen.M,pa=.3,pb=.5)

# without specifying allele frequncy, defaults to .5 for both loci
mod1 <- pen.to.stacd(pen.M)
mod2 <- pen.to.stacd(pen.M,pa=.5,pb=.5)
all.equal(mod1,mod2)
```

`stacd.data` *Generate data under STACD model*

Description

Generates data following the STACD model with the specified model parameters. The phenotype may be generated based on existing genotypes, or genotype data can be generated.

Usage

```
stacd.data(betas, x = NULL, z = NULL, n = NULL,
        ncase=NULL, ncontrol=NULL, pa = NULL, pb = NULL,
        model = "logistic", ve = NULL, verbose = FALSE)
```

Arguments

`betas`	a length 9 numeric vector of the STACD model parameters (in order: *mu*, *alpha_1*, *alpha_2*, *beta_1*, *beta2*, *gamma_ab*, *gamma^*_aab*, *gamma^*_abb*, and *gamma^*_aabb*)

x	a numeric vector of genotype data at the first locus, coded 0,1,2 for the number of reference alleles.
z	data for the second locus, as x
n	sample size for the generated data. Ignored with a warning if data is provided for x and/or z.
ncase	number of affected individuals to include in the generated data. Only applicable to the \"logistic\" model. Ignored with a warning if data is provided for x and/or z.
ncontrol	number of unaffected individuals to include in the generated data. Only applicable to the \"logistic\" model. Ignored with a warning if data is provided for x and/or z.
pa	allele frequncy for the first locus. Ignored with a warning if x is given.
pb	allele frequency for the second locus. Ignored with a warning if z is given.
model	either "normal" or "logistic", specifying the scale of the STACD model. If "normal", then the phenotype is continuous with normally distributed errors. If "logistic", then the phenotype is dichotomous based on an underlying continuous variable with logistically distributed errors.
ve	The variance of the error term in the normal model. If not specified, it will be set so that the total phenotypic variance is 1. If model="logistic", this argument is ignored (with a warning).
verbose	Logical, indicating whether to provide additional information about the model. Default is FALSE.

Details

Data may be generated based on existing genotype data (coded 0,1,2 or NA), or new genotype data may be generated from *Bin(2,p)*. If existing genotype data is provided for x and z then the sample size n and allele frequencies pa and pb are ignored. If no data is provided then the sample size and allele frequencies must be given so that new genotype data can be generated. For the logistic model, the sample size can optionally be specified to include a specific number of cases (y=1) and controls (y=0), instead of a simple random sample of size n from the population.

Optionally, by setting verbose=TRUE expected characteristics of the generated data will be provided in the format of stacd.model. If existing genotype data was used to generate the phenotype, these expected values and the variance decomposition are computed based on the observed allele frequencies in the input data.

The two loci for the genotype data are assumed to be biallelic, unlinked, and under Hardy-Weinberg equilibrium.

Value

If verbose=FALSE, the function returns a data frame with columns y, x, and z, where y is the generated phenotype data and x, and z are either the provided genotype data or new generated genotypes.

If verbose=TRUE, more complete information is returned. The output will include:

Full_Data	data frame containing the full generated data, including latent values where applicable. Specifically, y - the generated phenotype ystar - (logistic model only) the latent continuous variable dichotomize to y XB - the value of the linear model minus the error term yprob - (logistic model only) the probability y=1 conditional on the genotypes x - genotype data for the first locus z - genotype data for the second locus

`Model_Info` Expected values and variance decomposition for the model. If existing genotype data was used for `x` or `z`, model information is computed based on the observed allele frequency in the data. If a fixed number of cases and controls were generated, this information reflects the population from which the individuals were sampled, not expected sample statistics. See `stacd.model` for details on the provided information.

Author(s)

Raymond Walters

References

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

See Also

`stacd.model`

Examples

```
# Example from Walters et al:
exparams <- varex.to.stacd(varex=rep(0.05,8), pa=0.3, pb=0.3,
                           model="logistic", decomp="parameter",
                           mu=0, ve=0.6)
data5k <- stacd.data(betas=exparams$Betas, n=5000,
                     pa=0.3, pb=0.3, model="logistic")
head(data5k)

# generate data from existing genotypes
x <- rbinom(100,2,.5)
z <- rbinom(100,2,.3)
betas <- c(0,0,0,0,0,0,.25,0,.5)
dat.M <- stacd.data(betas,x,z)
head(dat.M)

# extra model output
mod3 <- stacd.data(betas,n=100,pa=.3,pb=.5,verbose=TRUE)
dat3.M <- mod3$Full_Data
head(dat3.M)
mod3$Model_Info
```

---

stacd.design *Create design matrix for STACD model*

Description

The STACD model is defined based on data for two unlinked biallelic loci. In order to fit the model, it is necessary to convert the data for the two loci into appropriate terms for the eight corresponding model parameters. This function creates that *N by 8* design matrix from the data.

Usage

```
stacd.design(x, z)
```

Arguments

x	the numeric vector of data at the first locus, coded 0,1,2 for the number of reference alleles
z	data for the second locus, as x

Value

An *N x 8* design matrix, with columns corresponding to the parameters of the STACD model (excluding *mu*), where N is the length of the data vectors x and z. Specifically, the columns of the output contain the appropriate dummy variables for *alpha_1*, *alpha_2*, *beta_1*, *beta_2*, *gamma_ab*, *gamma^*_aab*, *gamma^*_abb*, and *gamma^*_aabb*, respectively.

Author(s)

Raymond Walters

References

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

See Also

stacd.data to generate data for the STACD model

Examples

```
# Generate random data for 2 loci
x <- rbinom(100,2,.5)
z <- rbinom(100,2,.3)

# Create design matrix
dat.M <- stacd.design(x,z)
dat.M

# fit STACD model using design matrix
u <- .2*x + .2*(x*z)
y <- rbinom(1,1,u)
fit <- glm(y~.,data=cbind(y=y,dat.M),family=binomial("logit"))
summary(fit)
```

`stacd.model` *Information about STACD model*

Description

Based on the parameters for the STACD model it is possible to compute an array of information regarding the expected distribution of the phenotype. Specifically, we provide the conditional and marginal expected values for the phenotype, and the decomposition of the variance explained by the two loci.

Usage

```
stacd.model(betas, pa = NA, pb = NA, model = "logistic", ve = NA)
```

Arguments

`betas`	a length 9 numeric vector of the STACD model parameters (in order: *mu*, *alpha_1*, *alpha_2*, *beta_1*, *beta2*, *gamma_ab*, *gamma^*_aab*, *gamma^*_abb*, and *gamma^*_aabb*)
`pa`	(optional) allele frequency for the first locus. If omitted, defaults to .5 with a warning.
`pb`	(optional) allele frequency for the second locus. If omitted, defaults to .5 with a warning.
`model`	either "normal" or "logistic" specifying the scale of the STACD model. If "normal", then the phenotype is assumed to be continuous with normally distributed errors. If "logistic", then the phenotype is assumed to be dichotomous based on an underlying continuous variable with logistically distributed errors.
`ve`	the variance of the error term in the normal model. If not specified, it will be set so that the total phenotypic variance is 1. If model="logistic", this argument is ignored (with a warning).

Details

Conditional expected values and variance decomposition are computed based on the model parameters and the allele frequencies. If allele frequencies `pa` and `pb` are not given they are assumed to be 0.5 (with a warning). The two loci are assumed to be uncorrelated.

For the logistic model, the variance decomposition is based on the latent variable `y_star`, as is consistent with the pseudo-R^2 for logistic regression introduced by McKelvey & Zavoina (1975). The variable `y_star` is assume to be continuous with errors following a standard logistic distribution; the dichotmous phenotype y is created by thresholding this variable.

The two loci for the STACD model are assumed to be biallelic, unlinked, and under Hardy-Weinberg equilibrium.

Value

`Betas`	STACD model parameters

`Variance_Explained`
: a list with elements:

  `Multiple_R_Sq` - total proportion of variance explained by ther model

  `Parameter_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the STACD paramters

  `Cockerham_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the conventional decomposition given by Cockerham (1954).

In addition, information about the conditional expected values and error variance the phenotype is computed. The format for this information depends on whether the "normal" or "logistic" model is specified. For the logistic model, the ouput includes

`Penetrances` the provided 3x3 penetrance matrix

`Marginals` length 3 vectors of the marginal pentrance conditional on the first locus and on the second locus

`Prevalence` expected prevalence of the dichotmoous phenotype

`Var_ystar` variance of the continuous variable underlying the phenotype

For the normal model, the output instead includes

`Expected_Values`
: a list with elements:

  `Conditional` - a 3x3 matrix of the expected value of the phenotype by genotype

  `MarginA` - length 3 vector of the marginal expected value conditional on the genotype of the first locus

  `MarginB` - length 3 vector of the marginal expected value conditional on the genotype of the second locus

  `Unconditional` - overall expected value of the phenotype

`Var_y` expected variance of the phenotype

`error_variance`
: variance of the (normally-distrbuted) error term for the phenotype

Author(s)

Raymond Walters

References

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39: 859-882.

McKelvey, R., and Zavoina, W. (1975). A Statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103-120.

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

See Also

To solve for the STACD parameters based on other information, see `varex.to.stacd` and `pen.to.stacd`.

Examples

```
# define basic model information
betas <- c(0,0,0,0,0,0,.25,0,.5)
pa <- .3
pb <- .5

# get expected values, parameter decomposition
stacd.model(betas, pa, pb, model = "logistic")

# different output format for normal model
stacd.model(betas, pa, pb, model = "normal")
```

`stacd.power` *Power analysis for STACD model parameters*

Description

Performs power analysis for a single parameter in the STACD model as described in Walters et al. This power analysis is based on simulating data using the STACD model and computating of the Wald test for the parameter of interest. The power to detect the parameter is estimated by the proportion of observed significant results for the Wald test.

Usage

```
stacd.power(betas,param,N,nrep,
        pa,pb,model="logistic",ve=NA,
        alpha=.05,plotit=TRUE)
```

Arguments

`betas`	a length 9 numeric vector of the STACD model parameters (in order: *mu*, *alpha_1*, *alpha_2*, *beta_1*, *beta2*, *gamma_ab*, *gamma^*_aab*, *gamma^*_abb*, and *gamma^*_aabb*)
`param`	the name of the parameter to test power for; one of c("alpha_1", "alpha_2", "beta_1", "beta_2", "gamma_ab", "gamma^*_aab", "gamma^*_abb", or "gamma^*_aabb")
`N`	a numeric vector of which sample sizes to use for estimating power
`nrep`	the number of replications to simulate
`pa`	allele frequency for the first locus
`pb`	allele frequency for the second locus
`model`	either "logistic" (default) or "normal" specifying which version of the STACD model to use
`ve`	variance of the error term for the normal model
`alpha`	significance level for testing power
`plotit`	logical, if `TRUE` a plot of the estimated power at each sample size will be produced

Details

For each sample size N, data is generated according to the STACD model with the specified betas, and a Wald test is performed for the parameter of interest. This procedure is repeated for nrep replications, and the power at each sample size is estimated as the proportion of Wald tests that were observed to be significant at the given alpha level.

Value

A length(N) by 2 matrix of the estimated power at each given sample size. The estimate power is the proportion of nrep replications where the Wald test for the specified parameter was significant in the generated data. If plotit=TRUE, a line plot of the power as a function of sample size is also produced.

Note

Higher values of nrep will produce more accurate estimates of power, but may be slow.

Monte Carlo power estimation with stacd.power is intended for use with the logistic STACD model. For the normal model, Monte Carlo estimation is less useful since exact power under the model can be computed.

Author(s)

Charles Laurin, Raymond Walters

References

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

Examples

```
# Example from Walters et al.:
exvarex <- logodds.to.stacd(logodds.M=matrix(c(0,0,0, 0,0,.25, 0,0,1),3,3,byrow=TRUE),
                            pa=0.3, pb=0.3)
## Not run: stacd.power(betas=exvarex$Betas,param="gamma^*_aab",
    N=c(1250,2500,5000,10000,15000,20000,25000,27500,30000),
    pa=.3,pb=.3,nrep=100)
## End(Not run)

# Smaller example without plot
stacd.power(betas=exvarex$Betas,param="gamma^*_aab",
    N=c(500,1500,2500),nrep=100,pa=.3,pb=.3,plotit=FALSE)
```

---

varex.to.stacd *Convert variance decomposition to STACD parameters*

Description

Convert variance decomposition to STACD parameters

Usage

```
varex.to.stacd(varex, pa, pb, model = "logistic",
        decomp = "cockerham", mu = NA, ve = NA,
        sign = rep(1, 8))
```

Arguments

varex	an 8 element vector of non-negative values with sum < 1. The meaning of these values depends on decomp (see Details).
pa	allele frequency for the first locus
pb	allele frequency for the second locus
model	either "normal" or "logistic" specifying the scale of the STACD model. If "normal", then the phenotype is continuous with normally distributed errors. If "logistic", then the phenotype is dichotomous based on a latent continuous variable with logistically distributed errors.
decomp	a string specifying either the "parameter" or "cockerham" variance decomposition (see Details)
mu	The mu parameter for the STACD model. It does not affect the variance decomposition so can be specified separately. By default set so that E(y)=0 for the normal model or E(ystar)=0 for the logistic model.
ve	The variance of the error term in the normal model. If not specified for the normal model, ve will be set so that the total phenotypic variance is 1. This argument is ignored for the logistic model (with a warning).
sign	a length 8 vector with elements 1 or -1 denoting the sign of the square root to use with each variance component. By default, all positive square roots are used. See Details.

Details

The STACD model parameters are computed based on the provided variance decomposition and allele frequencies. For the "parameter" decomposition, the elements of varex are taken to be the semipartial correlation of the corresponding STACD parameter conditional on all lower order effects, i.e.

varex = *(R^2_alpha_1, R^2_alpha_2|alpha_1, R^2_beta_1, R^2_beta_2|beta_1, R^2_gamma_ab|alpha_1,alpha_2,bet*
R^2_gamma^*_aab|alpha_1,alpha_2,beta_1,beta_2,gamma_ab, R^2_gamma^*_abb|alpha_1,alpha_2,beta_1,beta_2,g
R^2_gamma^*_aabb|alpha_1,alpha_2,beta_1,beta_2,gamma_ab,gamma_aab^*,gamma_abb^*)

For the "cockerham" decomposition, the elements of varex instead correspond to the additive, dominant, and interaction variance components defined by Cockerham (1954), i.e.

varex = *(V_addA,V_domA,V_addB,V_domB, V_addxadd,V_addAxdomB,V_domAxaddB,V_domxdom)/Vary*

In either case, the the variance decomposition does not determine the value of mu, so it needs to be set separatly. For the logistic model, adjusting mu helps control the prevalence.

For the logistic model, the variance decomposition is applied following the approach of McKelvey & Zavoina to parition the variance of a continuous latent variable with logistically-distributed errors that is then dichotomized to the observed phenotype.

In addition to the STACD parameters, additional model information is provided using the same format as stacd.model. This information includes conditional and marginal expected values, and both variance decompositions.

Since the STACD parameters are a function of the square root of the variance components they are not uniquely defined. Instead it is necessary to decide whether to use the positive or the negative square root of each variance component. The `sign` argument is a vector with elements 1 or -1 to specify which square root to use for each variance component. By default, the positive root is used for all components.

Value

`Betas` STACD model parameters

`Variance_Explained`
: a list with elements:
  `Multiple_R_Sq` - total proportion of variance explained by ther model
  `Parameter_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the STACD paramters
  `Cockerham_Decomp` - a length 8 vector giving the decomposition of the proportion of variance explained based on the conventional decomposition given by Cockerham (1954).

In addition, information about the conditional expected values and error variance the phenotype is computed. The format for this information depends on whether the "normal" or "logistic" model is specified. For the logistic model, the ouput includes

`Penetrances` the provided 3x3 penetrance matrix

`Marginals` length 3 vectors of the marginal pentrance conditional on the first locus and on the second locus

`Prevalence` expected prevalence of the dichotmoous phenotype

`Var_ystar` variance of the continuous variable underlying the phenotype

For the normal model, the output instead includes

`Expected_Values`
: a list with elements:
  `Conditional` - a 3x3 matrix of the expected value of the phenotype by genotype
  `MarginA` - length 3 vector of the marginal expected value conditional on the genotype of the first locus
  `MarginB` - length 3 vector of the marginal expected value conditional on the genotype of the second locus
  `Unconditional` - overall expected value of the phenotype

`Var_y` expected variance of the phenotype

`error_variance`
: variance of the (normally-distrbuted) error term for the phenotype

Note

Since the STACD parameters are not uniquely defined by the variance components, the results of `varex.to.stacd` may not always exactly reverse `stacd.model`. In other words, if `stacd.model` is used to compute a variance decomposition, and `varex.to.stacd` is run with those variance components, `varex.to.stacd` is not guaranteed to reproduce the STACD parameters originally input into `stacd.model` (the STACD parameters can be reproduced with the correct setting for `sign`, but are not guaranteed by default). The variance components for the models will be equivalent, but the parmeters may differ.

Author(s)

Raymond Walters

References

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39: 859-882.

McKelvey, R., and Zavoina, W. (1975). A Statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103-120.

Walters, R., Laurin, C., and Lubke, G.H. (in prep). Derivation of a logistic model for two-locus genetic interactions.

See Also

For variance decomposition of given STACD parameters, see `stacd.model`

Examples

```
# Example from Walters et al
exparams <- varex.to.stacd(varex=rep(0.05,8), pa=0.3, pb=0.3,
                           model="logistic", decomp="parameter",
                           mu=0, ve=0.6)
print(exparams)

# additive effect of locus 1,
# additive x additive interaction, and
# dominant x dominant interaction
varex <- c(.003,0,0,0,.003,0,0,.003)
varex.to.stacd(varex, pa=.3, pb=.5)

# adjust prevalence
varex.to.stacd(varex, pa=.3, pb=.5, mu=-4)

# alternative parameters with same variance decomp
varex.to.stacd(varex, pa=.3, pb=.5, sign=c(rep(-1,4),rep(1,4)))

# alternative decomposition
varex.to.stacd(varex, pa=.3, pb=.5, decomp="parameter")
```

Index