**队列统计学脚本cohortcode.r封装函数使用手册V1.0**

撰写日期和版本：2024年08月23日

说明书撰写人：Yinglian Chen

脚本编写人：Yinglian Chen

概述：R脚本主要根据队列统计研究需求编写或收集相关R包使用方法，根据相似的内容封装或简化其使用过程，对使用参数接口进行标准化，便于软件开发过程中不同程序调用过程中的参数对接和简易使用，相关生信分析人员也可直接调用函数进行相关项目的数据分析，内容中不包含任何绘图内容，仅包含相关底层结果数据。

本说明书中不包括每种分析方法所解决的科学问题描述和结果解释，主要仅为程序和软件开发装配使用。

|  |  |  |
| --- | --- | --- |
| 表1：脚本中用到的程序包 | | |
| R包名称 | 使用功能 | 备注 |
| ade4 | PCoA | 用于主坐标转换分析 |
| vegan | PCoA | 用于主坐标转换分析 |
| poLCA | LCA | 用于潜在分类分析 |
| reshape2 | LCA | 用于潜在分类分析 |
| mclust | LPAs | 用于潜在剖面分析 |
| glmnet | chen\_lasso, chen\_rigde | 用于lasso和岭回归模型 |
| e1071 | chen\_svm | 用于支持向量机模型 |
| randomForest | chen\_randomforest | 用于随机森林模型 |
| pROC | modevalue | 用于模型评估 |
| survival | chen\_survival | 用于生存分析 |
| tableone | chen\_basetable | 用于基线表统计分分析 |

**函数目录**

[功能1. SNP位点遗传分析 4](#_Toc175231113)

[comg(datap) 4](#_Toc175231114)

[gld(x, y) 4](#_Toc175231115)

[mgld(mdata, i) 5](#_Toc175231116)

[功能2：数据降维 5](#_Toc175231117)

[PCA(mydata) 5](#_Toc175231118)

[PCoA(df, method, nf) 6](#_Toc175231119)

[ftsne(data) 6](#_Toc175231120)

[fumap(data, seed) 7](#_Toc175231121)

[功能3潜在分类和潜在剖面分析 8](#_Toc175231122)

[LCA(data, colns, nclass, maxiter=3000) 8](#_Toc175231123)

[LPAs(LPA\_data, colns, G, modelNames) 8](#_Toc175231124)

[功能4 模型分析 9](#_Toc175231125)

[chen\_lm 10](#_Toc175231126)

[chen\_glm 10](#_Toc175231127)

[chen\_lasso 10](#_Toc175231128)

[chen\_ridge 10](#_Toc175231129)

[chen\_svm 10](#_Toc175231130)

[chen\_randomforest 10](#_Toc175231131)

[功能5 模型评估 11](#_Toc175231132)

[modevalue(truevalues, predictions, ROC, method) 11](#_Toc175231133)

[功能6 生存分析 12](#_Toc175231134)

[chen\_survival (sdata, time, status, x) 12](#_Toc175231135)

[功能7 基线数据统计 12](#_Toc175231136)

[chen\_basetable(data, allvars, targetfactor, fvars, addOverall) 12](#_Toc175231137)

[测试文件目录 13](#_Toc175231138)

# 功能1. SNP位点遗传分析

## comg(datap)

功能注释：用于对独立单一的SNP位点基础信息计算，群体遗传信息评估。

输入变量：输入datap 类型为字符串向量，表示该遗传位点的基因型信息，多倍体使用逗号“，”隔开，单倍体信息也可按照相同方式输入运算。

输出变量：结果输出为list类型，包括：样本数目$samplenumber；基因型频率$GenetypeFrq；基因频率$GeneFrq；哈代平衡P值$HWEP；多态信息含量$PIC

位点杂合度$He；位点纯合度$Ho；有效等位基因数$Ne；位点固定世代数$Tfix；位点流失世代数$Tloss；近郊系数$FIS。

案例：

data=read.csv("1. 遗传数据分析测试样板.csv")

comg(data$rs2071449)

## gld(x, y)

功能注释：用于评估x和y变量两个SNP位点的连锁情况

输入变量：变量x和y与datap类型一致，为字符串向量，表示该遗传位点的基因型信息，多倍体使用逗号“，”隔开。

输出变量：结果输出为list类型，包括：两SNP位点信息$info；连锁不平衡D值 $D；标准化连锁不平衡D'值$`D'` ；位点相关度R2 $`R^2`。

案例：

data=read.csv("1. 遗传数据分析测试样板.csv")

gld(data$rs2071449,data$rs190771160)

## mgld(mdata, i)

功能注释：用于二维结构化数据举证中批量运算位点与位点之间的连锁不平衡，为gld(x,y)函数的输出矩阵版本。

输入变量：变量mdata为数据框类型，包括位点所有位点和对应人群信息

结果输出为matrix矩阵类型，包括每个位点之间连锁不平衡评估值；参数i可填写范围为2-4，默认为4，表示结果矩阵输出为R2，设置为2表示矩阵输出D值，3表示输出为标准化D'值。

输出变量：结果输出为matrix类型，为位点与位点之间的连锁值矩阵，值的类型由变量参数i决定。

案例：

data=read.csv("1. 遗传数据分析测试样板.csv")

mgld(data[,-1], 3)

# 功能2：数据降维

## PCA(mydata)

功能注释：用于多变量数据降维的主成分分析方法，将多个指标的信息降维转为2-3个指标信息，使样本在2或3维图形中可区分性。

输入变量：变量mydata类型为数据框，包括个条目对应指标变量信息。

输入变量：结果为list类型，包括：降维结果信息$layout；降维后指标$pc1，$pc2和$pc3所包含的方差占比信息。

案例：

data=read.csv("2. 降维分析 潜在剖面分析 模型分析 测试样板.csv")

PCA(data)

## PCoA(df, method, nf)

功能注释：类似PCA用于多变量数据降维的主坐标分析方法，将多个指标的信息降维转为2-3个指标信息，使样本在2或3维图形中可区分性。

输入变量：变量mydata类型为数据框，包括个条目对应指标变量信息。其他参数变量method参数为距离运算方法，默认为"bray"，距离方法包括："manhattan", "euclidean", "canberra", "bray", "kulczynski", "gower", "morisita", "horn", "mountford", "jaccard", "raup", "binomial", "chao", "altGower", "cao", "mahalanobis", "clark", "chisq", "chord", "hellinger", "aitchison"和"robust.aitchison"；nf参数为结果需要的维度设置，默认为2，3维图形展示可设置为3或更大值。

结果输出：为list类型包括：降维结果信息$layout；降维后指标$pc1，$pc2和$pc3所包含的代表性占比信息。

案例：

data=read.csv("2. 降维分析 潜在剖面分析 模型分析 测试样板.csv")

PCoA(data)

## ftsne(data)

功能注释：主要用于非正态分布或秩序多变量数据降维，将多个指标的信息降维转为2个指标信息，使样本在2维图形中可区分性，主要调用tsne包。

输入变量：data类型为数据框，包括个条目对应指标变量信息，其他参数变量主要参考tsne包，默认2维降维k=2，需要调整其他参数可直接调用tsne()函数，算法涉及随机抽样可设置种子函数set.seed()保证结果的可复现性。

输出变量：结果为list类型，主要结果为降维结果信息$layout。

案例：

data=read.csv("2. 降维分析 潜在剖面分析 模型分析 测试样板.csv")

ftsne(data)

## fumap(data, seed)

功能注释：主要用于非正态分布或秩序多变量数据降维，将多个指标的信息降维转为2个指标信息，使样本在2维图形中可区分性，主要调用umap包，相对与tsne来说map的方法不需要调用较大的运行内存，且运算速度较快。

输入变量：data类型为数据框，包括个条目对应指标变量信息；参数seed为种子函数，默认为123，算法涉及随机性，设置种子函数可保证结果的复现性。

输出变量：结果输出为list类型，包括：降维结果信息$layout；原始数据$data；

推荐分类数$knn；运行参数配置信息$config。

其他参数变量：

案例：

data=read.csv("2. 降维分析 潜在剖面分析 模型分析 测试样板.csv")

fumap(data)

# 功能3潜在分类和潜在剖面分析

## LCA(data, colns, nclass, maxiter=3000)

功能注释：在群体中根据群体多个指标特征，对群体进行潜在特征分类。

输入变量：data多个指标特征为非连续变量指标数据框类型，colns为data数据框中需要纳入分析指标的列名向量参数，nclass为整数型结果的分类数量设置，matiter指标默认为3000，用于运算或模型检验需求。

输出变量：结果为list类型结果，其中主要结果包括：AIC统计量$aic；

BIC统计量$bic；预测分类结果$predclass；预测的分类概率$posterior。

案例：

data=read.csv("3. 潜在分类分 模型分析 测试样板.csv")

LCA(data, colnames(data), 3)

## LPAs(LPA\_data, colns, G, modelNames)

功能注释：在群体中根据群体多个连续变量指标特征，对群体进行潜在剖面分类，主要使用mclust包调用。

输入变量：data多个指标特征为非连续变量指标数据框类型，colns为data数据框中需要纳入分析指标的列名向量参数，nclass为整数型结果的分类数量设置，

输入变量LPA\_data多个指标特征为连续变量指标数据框类型，colns为LPA\_data数据框中需要纳入分析指标的列名向量参数，G为整数型分类剖面数量设置，modelNames为使用剖面方法名称，默认为NULL，设置最优方法，可根据mlcuster调用其他剖面方法，包括"EII","VII","EEI","VEI","EVI","VVI","EEE","VEE","EVE VVE","EEV","VEV","EVV"和"VVV"方法，对于模型分析可使用PLA()或mclusterBIC()函数查看最优方法，对于个数选择可使用mclusterBootstrapLRT()函数进行评估。

输出变量：为list类型结果，其中主要结果包括：BIC统计量$BIC；ICL统计量$icl；预测剖面分类$classification；预测剖面分类概率$z。

案例：

data=read.csv("2. 降维分析 潜在剖面分析 模型分析 测试样板.csv")

LPAs(data, colnames(data), 3)

# 功能4 模型分析

功能描述：功能包括线性回归模型【chen\_lm】逻辑回归模型【chen\_glm】LASSO回归模型【chen\_lasso】岭回归模型【chen\_ridge】支持向量机SVM模型【chen\_svm】随机森林模型【chen\_randomforest】，模型输入调用基本一致，主要调用相关程序包使用，模型包括表2清单。

|  |  |  |
| --- | --- | --- |
| 表2：模型函数清单 | | |
| 函数名称 | 模型 | 调用包函数 |
| chen\_lm | 线性回归模型 | R自带函数 |
| chen\_glm | 逻辑回归模型 | R自带函数 |
| chen\_lasso | LASS回归模型 | R包glmnet |
| chen\_ridge | 岭回归模型 | R包glmnet |
| chen\_svm | 支持向量机模型 | R包e1071 |
| chen\_randomforest | 随机森林模型 | R包randomForest |

输入变量：模型同一输入类型，主要包括3个参数，1. data参数为数据框类型，包括需要分析的所有数据；2. x参数为字符向量类型，表示在data数据中的自变量组，以data数据参与自变量的列名，当data数据全非因变量列全参与时设置为点“.”；3. y参数为字符串类型，表示data数据中作为因变量的列名。

输出变量：模型结果，除支持向量机模型和随机森林模型外主要结果为模型系数矩阵$coef，向量机模型和随机森林模型保留$mod数据用于预测函数predict()使用或保存为二进制文件调用，线性模型保留F统计量、R相关性和模型P值等结果，逻辑回归模型保留AIC和BIC结果，随机森林保存参数重要性结果$importance。

案例：

data=read.csv("3. 潜在分类分 模型分析 测试样板.csv")

chen\_lm(data,c("x1","x2","x3","x4","x5"),"y")

chen\_glm(data,c("x1","x2","x3","x4","x5"),"y")

chen\_lasso(data,c("x1","x2","x3","x4","x5"),"y")

chen\_ridge(data,c("x1","x2","x3","x4","x5"),"y")

chen\_svm(data,c("x1","x2","x3","x4","x5"),"y")

chen\_randomforest(data,c("x1","x2","x3","x4","x5"),"y")

# 功能5 模型评估

## modevalue(truevalues, predictions, ROC, method)

功能描述：通过预测值和真实值的结果，评估模型准确性

输入变量：truevalues变量表示真实数据组，为向量类型；predictions变量表示通过模型预测的预测数据组，为向量类型；ROC是否为二分类特征，为逻辑类型，默认为False； method预测变量和真实变量相关性评估方法，默认为pearson，还包括spearman和kendall。

输出变量：结果包括：相关性值R2$r2；相关性P值$rp；平均方差$mse；标准平均差$rmse；平均差$mae；AUC面积$auc(分类变量设置ROC参数为Ture可生成)。

案例：

modevalue(sample(1:2,100,replace = T), sample(1:2,100,replace = T), ROC=T)

# 功能6 生存分析

## chen\_survival (sdata, time, status, x)

功能描述：用于生存分析运算，其中包括log-rank方法和cox方法，主要调用survival包使用简化。

输入变量：输入数据为数据框类型，必须要包括时间列、状态列和指标列，指标列，其中指标列可包含多列。

输出变量：输出结果包括：log-rank检测F值$logrankf；log-rank 检测P值$logrankp；cox模型数据$coxmod；cox系数表$coxtable；分类指标统计表$plotfit。

案例：

data=read.csv("4. 生存分析测试样板.csv")

chen\_survival(sdata=data, time="time", status="state", x=c("index1","index2"))

# 功能7 基线数据统计

## chen\_basetable(data, allvars, targetfactor, fvars, addOverall)

功能描述：对群体队列数据进行一键式生成统计基线结果，并对分类变变量进行分组统计和比较，多组样本进行多重比较，统计依据变量类型选择统计方法，主要调用tableone包分析。

输入变量：输入变量为数据框格式，数据框包括多个指标信息，可提前调整数据框每一列的数据类型，进行调用时将自动根据变量类型进行统计分析；参数中data为输入数据变量，数据框格式；allvars为需要统计的变量data数据列名，字符向量格式；targetfactor用于分组的变量data数据列名，字符向量格式；fvars在allvars中属于因素变量或非连续变量的data数据列名，字符向量格式；addOverall为是否整体统计设置，在targetfactor分组数量中不为1时可用，逻辑格式，默认为False。

输出变量：输出结果主要为三线表内容格式，可通过summary函数进行详细结果输出，可通过print函数打印结果。

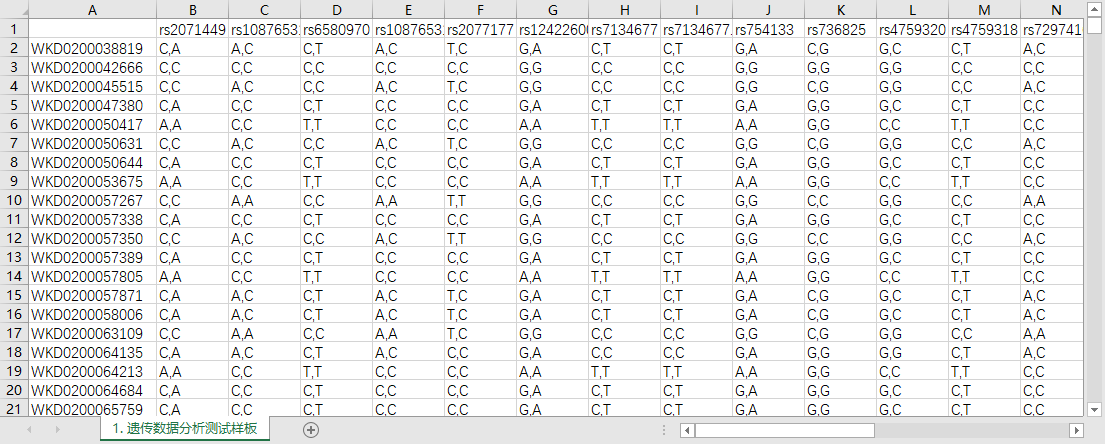
案例：

data=read.csv("4. 生存分析测试样板.csv")

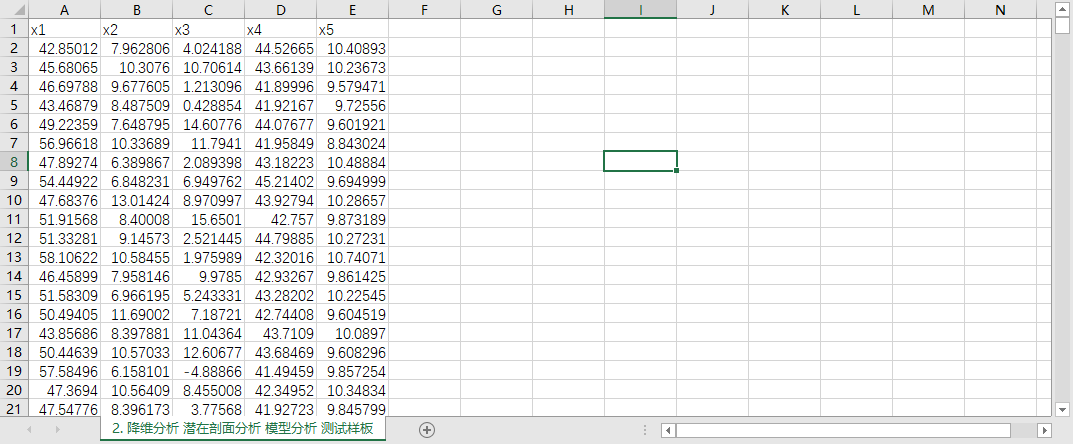
chen\_basetable(data=data, allvars=colnames(data), targetfactor="sex", fvars=c("state"), addOverall = T)

# 测试文件目录

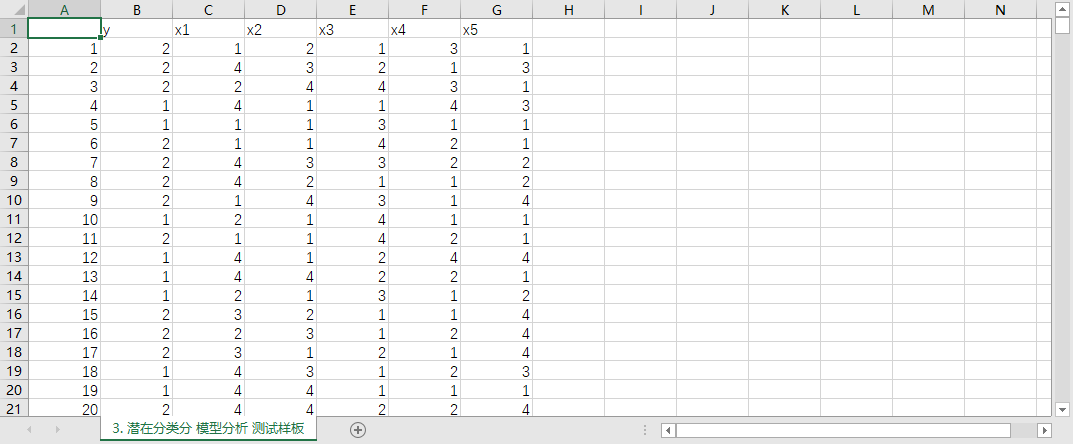
1. 遗传数据分析测试样板.csv



1. 降维分析 潜在剖面分析 模型分析 测试样板.csv



1. 潜在分类分 模型分析 测试样板.csv



4. 生存分析测试样板.csv

