Package 'SparseMSE'

March 4, 2019

Title Multiple Systems Estimation for Sparse Capture Data

Version 1.2.1

Author Lax Chan [aut, cre],
Bernard Silverman [aut],
Kyle Vincent [aut]

Maintainer Lax Chan <lax.chan1@nottingham.ac.uk>

Description Implements the routines and algorithms developed and analysed in ``Multiple systems estimation for Sparse Capture Data: Inferential Challenges when there are Non-Overlapping Lists" Chan, L, Silverman, B. W., Vincent, K (2019) <arXiv:1902.05156>. This package explicitly handles situations where there are pairs of lists which have no observed individuals in common.

URL https://arxiv.org/abs/1902.05156

Depends R (>= 3.5.0)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests rmarkdown

Imports lpSolve

NeedsCompilation no

Repository CRAN

Date/Publication 2019-03-04 16:40:03 UTC

R topics documented:

`Artificial_3` *Artificial data set to demonstrate possible instabilities*

Description

This is a simple data set based on three lists, which shows that there is not necessarily any clear hierarchical relationship between models that fail on one or the other of the criteria tested by `checkident`.

Usage

```
Artificial_3
```

Format

An object of class `data.frame` with 4 rows and 4 columns.

Details

If all three interactions are included in the fitted model then then the linear program in `checkident` yields a strictly positive value but the matrix A is not of full column rank, so the parameters are not identifiable. If the model contains AB either alone or in conjunction with one of AC and BC, then the linear program result is zero, so the MLE does not exist. If only main effects are considered, or if either or both of AC and BC, but not AB are included, then the model passes both tests.

`buildmodel` *Build model for multiple systems estimation*

Description

For multiple systems estimation model corresponding to a specified set of two-list effects, set up the GLM model formula and data matrix.

Usage

```
buildmodel(zdat, mX)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
mX	A $2 \times k$ matrix giving the k two-list interactions to be included in the model. Each column of mX contains the numbers of the corresponding pair of lists. If mX = 0, then all two-list interactions are included. If mX = NULL, no interactions are included and the main effects model is fitted.

Value

A list with components as below.

datamatrix A matrix with all possible capture histories, other than those corresponding to empty overlaps within the model. An empty overlap is a pair of lists (i, j) such that no case is observed in both lists, regardless of whether it is present on any other lists. If (i, j) is within the model specified by mX, all capture histories containing both i and j are then excluded.

modelform The model formula suitable to be called by the Generalized Linear Model function glm.

emptyoverlaps A matrix with two rows, whose columns give the indices of pairs of lists for which there are empty overlaps and where the pair is within the specified model. The column names give the names of the lists corresponding to each pair.

Examples

```
data(NewOrl)
buildmodel(NewOrl, mX=NULL)
```

---

buildmodelmatrix *Build the model matrix based on particular data, as required to check for identifiability and existence of the maximum likelihood estimate*

Description

This routine builds a model matrix as required by the linear program check checkident and checks if the matrix is of full rank. In addition, for each individual list, and for each pair of lists included in the model, it returns the total count of individuals appearing on the specific list or lists whether or not in combination with other lists.

Usage

```
buildmodelmatrix(zdat, mX = NULL)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
mX	A $2 \times k$ matrix giving the k two-list interactions to be included in the model. Each column of mX contains the numbers of the corresponding pair of lists. If mX = 0, then all two-list interactions are included. If mX = NULL, no interactions are included and the main effects model is fitted.

Value

A list with components as below

modmat The matrix that maps the parameters in the model (excluding any corresponding to non-overlapping lists) to the log expected value of the counts of capture histories that do not contain non-overlapping pairs in the data.

tvec A vector indexed by the parameters in the model, excluding those corresponding to non-overlapping pairs of lists. For each parameter the vector contains the total count of individuals in all the capture histories that dominate that parameter.

rankdef The column rank deficiency of the matrix modmat. If rankdef = 0, the matrix has full column rank.

Examples

```
data(NewOrl)
buildmodelmatrix(NewOrl, mX=NULL)
```

---

checkallmodels *Check all possible models*

Description

This routine checks every possible model for existence and identifiability of the maximum likelihood estimates.

Usage

```
checkallmodels(zdata, verbose = F)
```

Arguments

zdata	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
verbose	Specifies whether all possible models are listed in the output, or only those which generate a non-zero error code.

Details

The routine calls the routine checkident for every model. If there are t lists then there are $t(t-1)/2$ pairs of lists, and hence $2^{t(t-1)/2}$ possible models, because the models correspond to subsets of the set of all pairs of lists. If $t = 7$ there are 2,097,152 models to check, which would take several hours. If t is equal to 8 or more, the routine terminates with a statement of the number of models and an explanation that checking all of these is not possible in a reasonable time.

Value

If verbose=F, it gives a matrix each of whose rows specifies a model, with last entry equal to the error code. Only those models yielding a non-zero error code are included, so if no models lead to an error the matrix is empty. Each of the first $t(t-1)/2$ columns corresponds to a pair of lists, and for each row, presence in or absence from the corresponding model is indicated by the value 1 or 0 respectively.

If verbose=T, it gives the full matrix of models together with the error codes they generate.

Examples

```
data(Artificial_3)
data(Western)
checkallmodels(Artificial_3, verbose= TRUE)
checkallmodels(Western)
```

---

checkident — *Check a model for the existence and identifiability of the maximum likelihood estimate*

Description

Apply the linear programming test as derived by Fienberg and Rinaldo (2012), and a calculation of the rank of the design matrix, to check whether a particular model yields an identifiable maximum likelihood estimate based on the given data. The particular algorithm applied is described on page 3 of the supplementary material, with a typographical error corrected.

Usage

```
checkident(zdat, mX = 0, verbose = F)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
mX	A $2 \times k$ matrix giving the k two-list interactions to be included in the model. Each column of mX contains the numbers of the corresponding pair of lists. If mX = 0, then all two-list interactions are included. If mX = NULL, no interactions are included and the main effects model is fitted.
verbose	Specifies the output. If F then the error code is returned. If T then in addition the routine prints an error message if the model/data fail either of the two tests, and also returns both the error code and the lp object.

Value

If verbose=F, then return the error code ierr which is 1 if the linear program test shows that the maximum likelihood estimate does not exist, 2 if it is not identifiable, and 3 if both tests are failed.

If verbose=T, then return a list with components as below

ierr As described above.

zlp Linear programming object, in particular giving the value of the objective function at optimum.

References

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. Ann. Statist. 40, 996-1023. Supplementary material: Technical report, Carnegie Mellon University. Available from http://www.stat.cmu.edu/~arinaldo/Fienberg_Rinaldo_Supplementary_Material.pdf.

---

estimatepopulation *Estimate the total population including the dark figure*

Description

This routine calculates the estimate of the total population, including the dark figure, together with confidence intervals as specified. It also returns the details of the fitted model.

Usage

```
estimatepopulation(zdat, method = "stepwise", quantiles = c(0.025,
  0.975), mX = NULL, pthresh = 0.001)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
method	If `method = "stepwise"` the stepwise method implemented in `stepwisefit` is used. If `method = "fixed"` then a specified fixed model is used; the model is then given by `mX`. If `method = "main"` then main effects only are fitted.
quantiles	Quantiles of interest for confidence intervals.
mX	A $2 \times k$ matrix giving the k two-list interactions to be included in the model if `method = "fixed"`. Each column of `mX` contains the numbers of the corresponding pair of lists. If `mX = 0`, then all two-list interactions are included. If `mX = NULL`, no interactions are included and the main effects model is fitted. If only one interaction is to be fitted, it is ok to specify it as a vector of length 2, e.g `mX=c(1,3)` for interactions of list 1 and 3. If `method` is equal to `"stepwise"` or `"fixed"` then `mX` is ignored.
pthresh	Threshold p-value used if `method = "stepwise"`.

Value

A list of components as below

`estimate` Point estimate and confidence interval estimates corresponding to specified quantiles.

`MSEfit` The model fitted to the data in the format described in `modelfit`.

Examples

```
data(NewOrl)
data(NewOrl_5)
estimatepopulation(NewOrl, method="stepwise", quantiles=c(0.025,0.975))
estimatepopulation(NewOrl_5, method="main", quantiles=c(0.01, 0.05,0.95, 0.99))
```

`investigateAIC` *Plot of simulation study*

Description

This routine produces Figure 1 of Chan, Silverman and Vincent (2019).

Usage

```
investigateAIC(nsim = 10000, Nsamp = 1000, seed = 1001)
```

Arguments

nsim	The number of simulation replications
Nsamp	The expected value of the total population within each simulation
seed	The random number seed

Details

Simulations are carried out for two different three-list models. In one model, the probabilities of capture are 0.01, 0.04 and 0.2 for the three lists respectively, while in the other the probability is 0.3 on all three lists. In both cases, there are no interaction effects, so that captures on the lists occur independently of each other. The first model is chosen to be somewhat more typical of the sparse capture case, of the kind which often occurs in the human trafficking context, while the second is a more classical multiple systems estimate.

The probability of an individual having each possible capture history is first evaluated. Then these probabilities are multiplied by `Nsamp = 1000` and, for each simulation replicate, Poisson random values with expectations equal to these values are generated to give a full set of observed capture histories; together with the null capture history the expected number of counts (population size) is equal to `Nsamp`. Inference was carried out both for the model with main effects only, and for the model with the addition of an interaction effect between the first two lists. The reduction in deviance between the two models was determined.

Checking for compliance with the conditions for existence and identifiability of the estimates shows that a very small number of the simulations for the sparse model (two out of ten thousand) fail the checks for existence even within the extended maximum likelihood context. Detailed investigation shows that in neither of these cases is the dark figure itself not estimable; although the parameters themselves cannot all be estimated, there is a maximum likelihood estimate of the expected capture frequencies, and hence the deviance can still be calculated.

The routine produces QQ-plots of the resulting deviance reductions against quantiles of the χ^2_1 distribution, for `nsim` simulation replications.

Value

An `nsim` $\times$ 2 matrix giving the changes in deviance for each replication for each of the two models.

References

Chan, L., Silverman, B. W., and Vincent, K. (2019). Multiple systems estimation for Sparse Capture Data: Inferential Challenges when there are Non-Overlapping Lists. Available from https://arxiv.org/abs/1902.05156.

`modelfit` *Fit a specified model to Multiple Systems Estimation data*

Description

This routine fits a specified model to multiple systems estimation data, taking account of the possibility of empty overlaps between observed lists.

Usage

```
modelfit(zdat, mX = NULL, check = T)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
mX	A $2 \times k$ matrix giving the k two-list interactions to be included in the model. Each column of mX contains the numbers of the corresponding pair of lists. If mX = 0, then all two-list interactions are included. If mX = NULL, no interactions are included and the main effects model is fitted. If only one interaction is to be fitted, it may be specified as a vector of length 2, e.g mX=c(1,3) for interactions of list 1 and 3.
check	If check = T check first of all if the maximum likelihood estimate exists and is identifiable, using the routine checkident. If either condition fails, print an appropriate error message and return the error code.

Value

A list of components as below

fit Details of the fit of the specified model as output by glm. The Akaike information criterion is adjusted to take account of the number of parameters corresponding to empty overlaps.

emptyoverlaps Matrix with two rows, giving the list pairs within the model for which no cases are observed in common. Each column gives the indices of a pair of lists, with the names of the lists in the column name.

poisspempty the Poisson p-values of the empty overlaps.

Examples

```
data(NewOrl)
modelfit(NewOrl,c(1,3), check=TRUE)
```

NewOrl *New Orleans data*

Description

Victims related to modern slavery and human trafficking in New Orleans

Usage

```
NewOrl
```

Format

An object of class `data.frame` with 19 rows and 9 columns.

Details

These data are collected into 8 lists. For reasons of confidentiality the labels of the lists are anonymised. Fuller details are given in Bales, Murphy and Silverman (2018).

References

K. Bales, L. Murphy and B. W. Silverman (2018). How many trafficked and enslaved people are there in New Orleans? Available from https://tinyurl.com/ybfb9tg6.

NewOrl_5 *New Orleans data five list version*

Description

New Orleans data consolidated into five lists

Usage

```
NewOrl_5
```

Format

An object of class `data.frame` with 14 rows and 6 columns.

Details

This reduces the New Orleans data `NewOrl` into five lists, constructed by combining the four smallest lists B, E, F and G into a single list.

---

stepwisefit *Stepwise fit using Poisson pvalues.*

Description

Starting with a model with main effects only, pairwise interactions are added one by one. At each stage the interaction with the lowest p-value is added, provided that p-value is lower than `pthresh`, and provided that the resulting model does not fail either of the tests in `checkident`.

Usage

```
stepwisefit(zdat, pthresh = 0.001)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
pthresh	this is the threshold below which the p-value of the newly added parameter needs to be in order to be included in the model. If `pthresh = 0` then the model with main effects only is returned.

Details

For each candidate interaction for possible addition to the model, the p-value is calculated as follows. The total of cases occurring on both lists indexed by the interaction (regardless of whether or not they are on any other lists) is calculated. On the null hypothesis that the effect is not included in the model, this statistic has a Poisson distribution whose mean depends on the parameters within the model. The one-sided Poisson p-value of the observed statistic is calculated.

Value

A list with components as follow

`fit` Details of the fit of the specified model as output by `glm`. The Akaike information criterion is adjusted to take account of the parameters corresponding to empty overlaps.

`emptyoverlaps` Matrix with two rows, each column of which gives the list pairs within the model for which empty overlaps are observed.

`poisspempty` the Poisson p-value of the empty overlaps.

Examples

```
data(NewOrl)
stepwisefit(NewOrl, pthresh=0.001)
```

---

tidylists *Produce a data matrix with a unique row for each capture history*

Description

This routine finds rows with the same capture history and consolidates them into a single row whose count is the sum of counts of the relevant rows. If `includezerocounts = T` then it also includes rows for all the capture histories with zero counts; otherwise these are all removed.

Usage

```
tidylists(zdat, includezerocounts = F)
```

Arguments

zdat	Data matrix with $t + 1$ columns. The first t columns, each corresponding to a particular list, are 0s and 1s defining the capture histories observed. The last column is the count of cases with that particular capture history. List names A, B, ... are constructed if not supplied. Where a capture history is not explicitly listed, it is assumed that it has observed count zero.
includezerocounts	If F then remove rows corresponding to capture histories with zero count. If T then include all possible capture histories including those with zero count, excluding the all-zero row corresponding to the dark figure.

Value

A data matrix in the form specified above, including all capture histories with zero counts if includezerocounts=T.

Examples

```
data(NewOrl)
zdat<-tidylists(NewOrl,includezerocounts=TRUE)
```

Western *Victims related to sex trafficking in a U.S. Western site*

Description

These data are collected into 5 lists. For reasons of confidentiality the labels of the lists are anonymised. Fuller details are given in Farrell, Dank, Kfafian, Lockwood, Pfeffer, Hughes, Vincent (2019).

Usage

```
Western
```

Format

An object of class data.frame with 13 rows and 6 columns.

References

Farrell, A., Dank, M., Kfafian, M., Lockwood, S., Pfeffer, R., Hughes, A., and Vincent, K. (2019). Capturing human trafficking victimization through crime reporting. Technical Report 2015-VF-GX-0105, National Institute of Justice. Available from https://www.ncjrs.gov/pdffiles1/nij/grants/252520.pdf.

Index