

Package ‘mixreg’

November 4, 2009

Version 0.0-3

Date 2009-11-04

Title Functions to fit mixtures of regressions.

Author Rolf Turner <r.turner@auckland.ac.nz>

Maintainer Rolf Turner <r.turner@auckland.ac.nz>

Depends R (>= 0.99)

Description Fits mixtures of (possibly multivariate) regressions (which has been described as doing ANCOVA when you don't know the levels).

License GPL (>= 2)

URL <http://www.math.unb.ca/~rolf/>

Repository CRAN

Date/Publication 2009-11-04 07:53:25

R topics documented:

| | |
|-----------------------|----|
| aphids | 2 |
| bootcomp | 3 |
| cband | 5 |
| covmix | 7 |
| mixreg | 8 |
| plot.cband | 10 |
| plot.mresid | 11 |
| qq.mix | 12 |
| resid.mix | 13 |

| | |
|--------------|-----------|
| Index | 15 |
|--------------|-----------|

| | |
|--------|---|
| aphids | <i>Data on the rate of infection of tobacco plants by a virus spread by aphids.</i> |
|--------|---|

Description

The `aphids` data frame has 51 rows and 2 columns. These correspond to 51 independent experiments in which varying numbers of aphids were released in a flight chamber containing 12 infected and 69 healthy tobacco plants. The resulting number of infected plants was recorded.

Usage

```
data(aphids)
```

Format

This data frame contains the following columns:

n.aphids the number of aphids released in the flight chamber in each instance

n.inf the resulting number (out of a possible 69) of infected plants

DETERMINATION OF INFECTION

After 24 hours, the flight chamber was fumigated to kill the aphids, and the previously healthy plants were moved to a greenhouse and monitored to detect symptoms of infection. The number of plants displaying such symptoms was recorded.

Source

These data appear courtesy of Gilles Boiteau and George Tai of the Potato Research Centre, Agriculture and Agri-Food Canada, Fredericton, New Brunswick. Any published work using these data should cite the paper given in the **References**.

References

Boiteau, G., M. Singh, R. P. Singh, G. C. C. Tai, and T. R. Turner (1998). Rate of spread of PVY-n by alate *Myzus persicae* (Sulzer) from infected to healthy plants under laboratory conditions. *Potato Research*, vol. 41, pp. 335 – 344.

| | |
|----------|---|
| bootcomp | <i>Perform a bootstrap test for the number of components in a mixture of regressions.</i> |
|----------|---|

Description

Produces nboot bootstrap realizations of the likelihood ratio statistic, either parametrically or semi-parametrically, and calculates the corresponding p-value of the test.

Usage

```
bootcomp(x, y, ncomp=2, ncincr=1, intercept=TRUE, nboot=1000,
         ts1=NULL, ts2=NULL, sem.par=FALSE, verb=FALSE,
         print.prog=TRUE, ...)
```

Arguments

| | |
|------------|--|
| x | A matrix of predictors for each of the regression models in the mixture. It should NOT include an initial column of 1s. If there is only one predictor, x may be a vector. |
| y | The vector of responses for the regression models. |
| ncomp | The null-hypothesized number of components in the mixture. |
| ncincr | The increment from the null-hypothesized number of components in the mixture to the number under the alternative hypothesis; i.e. the number of components under the alternative hypothesis is ncomp + ncincr. |
| intercept | Logical argument indicating whether the regression models in the mixture should have intercept terms. |
| nboot | The number of bootstrap replicates of the log likelihood ratio statistic to be produced. |
| ts1 | Starting values for fitting the ncomp component model. If ts1 is null, random starting values are used. (This is not recommended.) |
| ts2 | Starting values for fitting the ncomp+ncincr component model. If ts2 is null, random starting values are used. (This is not recommended.) |
| sem.par | Logical argument indicating whether semi-parametric bootstrapping should be used. |
| verb | Logical argument indicating whether the fitting processes should be verbose (i.e. whether details should be printed out at each step of the EM algorithm). If TRUE a huge amount of screen output is produced. |
| print.prog | Logical argument indicating whether the index of the bootstrap replicate just completed should be printed out, to give an idea of how the process is progressing. |
| ... | Further arguments to be passed to mixreg to control the fitting procedure. |

Details

In parametric bootstrapping the bootstrap data sets are generated by simulating from the fitted `ncomp` model parameters, using Gaussian errors. In semi-parametric bootstrapping the errors are generated by resampling from the residuals. Since at each predictor vector there are `ncomp` residuals, one for each component of the model, the errors are selected from these `ncomp` possibilities. The selection probabilities at this step are the conditional probabilities, of the observation being generated by each component of the model, given that observation. These probabilities depend on the parameters of the model whence the procedure is semi-parametric.

Value

A list (of class "mixreg") with components

| | |
|-------------------------------|---|
| <code>lrs</code> | The log likelihood ratio statistic for testing that the number of components is <code>ncomp</code> versus that it is <code>ncomp + nincr</code> . |
| <code>aic.ncomp</code> | The vector (with dimension <code>nboot</code>) of Akaike Information Criterion values for each of the fitted <code>ncomp</code> component models fitted to bootstrap data sets. The value of <code>ncomp</code> is substituted in the name; e.g. if <code>ncomp = 2</code> then the name of this component of the returned list is "aic.2". |
| <code>aic.ncomp+ncincr</code> | The vector (with dimension <code>nboot</code>) of Akaike Information Criterion values for each of the fitted <code>ncomp+ncincr</code> component models fitted to bootstrap data sets. The value of <code>ncomp+ncincr</code> is substituted in the name; e.g. if <code>ncomp = 2</code> and <code>ncincr=1</code> , then the name of this component of the returned list is "aic.3". |
| <code>pval.boot</code> | The p-value of the hypothesis test from the bootstrapping procedure. It is calculated as $\text{sum}(lrs \leq lrs.boot)/nboot$. |
| <code>lrs.boot</code> | The vector of bootstrap replicates of the log likelihood ratio statistic |
| <code>screw.ups</code> | A list giving information about the screw-ups that have occurred in the bootstrapping procedure; it includes the values of <code>.Random.seed</code> that lead to the data causing the screw-up so that the difficulty may be re-produced and examined if so desired. See the comments in the code for the meaning of the various "types" of screw-up. The "times" component of the <code>screw.ups</code> list gives the index of the bootstrap replicate that was being worked on when the screw-up occurred. Note that if a screw-up does occur, the replicate is redone completely. |

References

Turner, T. R. Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. Submitted for publication, 1998.

See Also

[cband](#), [covmix](#), [mixreg](#), [plot.cband](#), [plot.mresid](#), [qq.mix](#), [resid.mix](#)

Examples

```

TS1 <- list(list(beta=c(3.0,0.1), sigsq=16, lambda=0.5),
            list(beta=c(0.0,0.0), sigsq=16, lambda=0.5))
TS2 <- list(list(beta=c(3.0,0.1), sigsq=9, lambda=1/3),
            list(beta=c(1.5,0.05), sigsq=9, lambda=1/3),
            list(beta=c(0.0,0.0), sigsq=9, lambda=1/3))
data(aphids)
x <- aphids$n.aphids
y <- aphids$n.inf
## Not run:
  nboot <- 1000
## End(Not run)

boot.23 <- bootcomp(x,y,nboot=nboot,ts1=TS1,ts2=TS2)

```

| | |
|-------|--|
| cband | <i>Calculate confidence and prediction bands for mixtures of one-variable regressions.</i> |
|-------|--|

Description

Produces confidence and prediction bands, two-sided or upper or lower, for the lines fitted in a model consisting of a mixture of one-variable regressions.

Usage

```
cband(object, cov.mat, x, y, alpha=0.05, xlen=100, plotit=FALSE,
      type=NULL)
```

Arguments

| | |
|---------|---|
| object | Object describing the fitted mixture of regressions, as returned by mixreg. |
| cov.mat | The estimated covariance matrix of the parameter estimates of the fit, as returned by covmix. |
| x | The vector of predictors for each of the regression models in the mixture. Only one-dimensional predictors are permitted here. |
| y | The vector of responses for the regression models. |
| alpha | One minus the confidence level for the confidence and prediction bands; e.g. alpha = 0.05 for 95% confidence. |
| xlen | The number of points to be plotted in the band envelopes. The x-values of the points will be equi-spaced from min(x) to max(x). |
| plotit | Logical argument indicating whether to plot the fitted model and confidence bands. |
| type | Character argument specifying the type of band. Must have one of the values "both", "upper", or "lower". Defaults to "both". |

Details

The prediction bands are conditional in that the associated probability is conditional upon the associated observation being generated by the relevant component of the mixture.

Value

A list (of class "cband") with components

| | |
|------------------------|---|
| <code>theta</code> | The parameter list from object (as returned by <code>mixreg</code>). |
| <code>intercept</code> | The logical value from object indicating whether intercepts were fitted. |
| <code>x</code> | The argument <code>x</code> of the call to <code>cband</code> . |
| <code>y</code> | The argument <code>y</code> of the call to <code>cband</code> . |
| <code>xf</code> | The equispaced sequence of values from <code>min(x)</code> to <code>max(x)</code> at which the values of the band envelopes were calculated. |
| <code>bnds</code> | A list with one entry for each component of the mixture. Each entry is a matrix with 4 columns (lower and upper confidence, lower and upper prediction bounds) if type is "both", or with 2 columns (confidence bounds, prediction bounds) if type is "upper" or "lower". |

Side Effects

If `plotit` is `TRUE` a plot of the fit and the confidence and prediction bands is produced in whatever device is currently open.

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Statist.* vol. 49, Part 3, pp. 371 – 384.

See Also

[bootcomp](#), [covmix](#), [mixreg](#), [plot.cband](#), [plot.mresid](#), [qq.mix](#), [resid.mix](#)

Examples

```
#See mixreg for examples.
```

| | |
|--------|---|
| covmix | <i>Calculate the covariance matrix of the parameter estimates for a mixture of regressions.</i> |
|--------|---|

Description

Produces an estimate of the covariance matrix of the parameter estimates for a fitted mixture of linear regressions, by inverting the observed Fisher information matrix.

Usage

```
covmix(object, x, y)
```

Arguments

| | |
|--------|--|
| object | Object describing the fitted mixture of regressions, as returned by mixreg. |
| x | A matrix of predictors for each of the regression models in the mixture. It should NOT include an initial column of 1s. If there is only one predictor, x may be a vector. |
| y | The vector of responses for the regression models. |

Details

If different variances are allowed amongst the components, the parameters are taken in the order $\beta_{.1}, \text{sig}^2_{.1}, \lambda_{.1}, \dots, \beta_{.K}, \text{sig}^2_{.K}$ for a K component model — $\lambda_{.K}$ is redundant and hence omitted. If equal variances are assumed, the parameters are taken in the order $\beta_{.1}, \lambda_{.1}, \dots, \beta_{.K}, \text{sig}^2$.

In the foregoing β refers to the linear coefficients, sig^2 to the variance, and λ to the mixing probability.

Value

The estimated covariance matrix.

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Statist.* vol. 49, Part 3, pp. 371 – 384.

Louis, T. A. Finding the observed information matrix when using the EM algorithm, *J. R. Statist. Soc. B*, vol. 44, pp. 226 – 233, 1982.

See Also

[bootcomp](#), [cband](#), [mixreg](#), [plot.cband](#), [plot.mresid](#), [qq.mix](#), [resid.mix](#)

Examples

```
#See mixreg for examples.
```

 mixreg

Fit a mixture of linear regressions.

Description

Estimates the parameters for a mixture of linear regressions, assuming Gaussian errors, using the EM algorithm.

Usage

```
mixreg(x, y, ncomp=2, intercept=TRUE, eq.var=FALSE,
       theta.start=NULL, itmax=1000, eps=1e-06, verb=TRUE,
       digits=7, max.try=5, data.name=NULL)
```

Arguments

| | |
|-------------|---|
| x | A matrix of predictors for each of the regression models in the mixture. It should NOT include an initial column of 1s. If there is only one predictor, x may be a vector. |
| y | The vector of responses for the regression models. |
| ncomp | The number of components in the mixture. |
| intercept | Logical argument specifying whether the linear regressions should have intercepts fitted. |
| eq.var | Logical argument specifying whether the error variance should be the same for all components, or each component should be allowed a different error variance. |
| theta.start | A list giving starting values for the estimation procedure. Each component of the list is in turn a list with components beta (vector of linear coefficients), sigsq (variance) and lambda (mixing probability). If eq.var is TRUE, then it is sensible to have all the starting values of sigsq equal, but this is not strictly necessary. If theta.start is not specified, starting values are generated randomly. This is NOT recommended. |
| itmax | The maximum number of EM steps to be undertaken. |
| eps | A value specifying the convergence criterion for the EM algorithm. If the maximum absolute value of the change in the parameters is less than eps the algorithm is considered to have converged. |
| verb | Logical argument; if verb is TRUE then details of the progress of the algorithm are printed out at each EM step. |
| digits | The number of digits to which the details are printed out, when verb is TRUE. |
| max.try | If the algorithm encounters a singularity in the likelihood (as may occur when eq.var is FALSE) the algorithm is restarted using new (randomly generated) starting values. The restart is attempted a maximum of max.try times. |
| data.name | A character string specifying a name associated with the data being analyzed, for identification purposes. |

Details

Even if `eq.var` is `TRUE`, each component of `theta` still has its own `sigsq` component. The values of these will all be equal however if `eq.var` is `TRUE`.

Value

A list, of class `mixreg`, with components

| | |
|------------------------|--|
| <code>parmat</code> | The parameters of the fitted model arranged as a matrix, each row corresponding to one component of the mixture. |
| <code>theta</code> | The parameters of the fitted model as a list, each entry of the list being itself a list (like those in <code>theta.start</code>) corresponding to one component of the mixture. |
| <code>log.like</code> | The log likelihood of the fitted model, based on Gaussian errors. |
| <code>aic</code> | The Akaike Information Criterion value for the fitted model; <code>aic</code> is equal to $-2 * \log.like + 2 * M$ where <code>M</code> is the number of parameters in the model. |
| <code>intercept</code> | The <code>intercept</code> argument of the call to <code>mixreg</code> . |
| <code>eq.var</code> | The <code>eq.var</code> argument of the call to <code>mixreg</code> . |
| <code>bnms</code> | A vector of names associated with the linear components of the regression models. The names are formed from the column names of the argument <code>x</code> if these exist; otherwise they are "beta1", "beta2", The name "Int" is prepended if <code>intercept</code> is <code>TRUE</code> . |
| <code>nsteps</code> | The number of steps the EM algorithm took to converge. |
| <code>converged</code> | Logical value indicating whether the algorithm did converge or stopped because it reach the <code>itmax</code> EM step. |
| <code>data.name</code> | The <code>data.name</code> argument if supplied; otherwise is formed as "name-of-y.on.name-of-x". |

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Statist.* vol. 49, Part 3, pp. 371 – 384.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statist. Soc. B*, vol. 39, pp. 1–22, 1977.

See Also

[bootcomp](#), [cband](#), [covmix](#), [plot.cband](#), [plot.mresid](#), [qq.mix](#), [resid.mix](#)

Examples

```
data(aphids)
x <- aphids$n.aphids
y <- aphids$n.inf
TS <- list(list(beta=c(3.0, 0.1), sigsq=16, lambda=0.5),
           list(beta=c(0.0, 0.0), sigsq=16, lambda=0.5))
fit <- mixreg(x, y, ncomp=2, theta.start=TS, data.name='aphids')
```

```

cvm <- covmix(fit,x,y)
cbd <- cband(fit,cvm,x,y)
plot(cbd)
r <- resid.mix(fit,x,y)
plot(r)
r <- resid.mix(fit,x,y,std=TRUE)
qq.mix(r)

```

plot.cband

Plot confidence bands for a mixture of regressions.

Description

Plots the fitted lines and confidence and prediction bands as calculated by cband, for a mixture of regressions on one variable.

Usage

```

## S3 method for class 'cband':
plot(x, cbands=TRUE, pbands=TRUE, xlab=NULL, ylab=NULL,
      main=NULL, ...)

```

Arguments

| | |
|--------|---|
| x | An object specifying the fitted lines and confidence and prediction bands, as produced by cband. |
| cbands | Logical argument specifying whether to plot the confidence bands. |
| pbands | Logical argument specifying whether to plot the prediction bands. |
| xlab | Character string specifying a label for the x-axis; defaults to "x". |
| ylab | Character string specifying a label for the y-axis; defaults to "y". |
| main | Character string specifying a title for the plot; if it is not specified a default title is formed. If no title at all is desired, specify main="". |
| ... | Optional extra arguments; not currently used. |

Details

This function is a "method" for plot. Note that a simple plot of the fit may be produced by specifying both cbands=FALSE and pbands=FALSE.

Side Effects

A plot is produced in whatever device is currently open.

See Also

[bootcomp](#), [cband](#), [covmix](#), [mixreg](#), [plot.mresid](#), [qq.mix](#), [resid.mix](#)

Examples

```
#See mixreg for examples.
```

```
plot.mresid          Plot residuals for a fitted mixture of linear regressions.
```

Description

Plots the residuals against predictors or fitted values using symbols whose size is proportional to the probability that the associated observation was generated by the associated component of the model.

Usage

```
## S3 method for class 'mresid':
plot(x, vs.fit=FALSE, whichx=1, shape="disc",
      ngon=20, size=1, xlab=NULL, ...)
```

Arguments

| | |
|---------------------|--|
| <code>x</code> | A list with components containing the residuals, the relevant probabilities, the predictors, and the observations, as returned by <code>resid.mix()</code> . When plotting against fitted values, it is probably better to use the standardized residuals, i.e. <code>resid.mix()</code> should be called with <code>std=TRUE</code> . |
| <code>vs.fit</code> | Logical argument. If <code>TRUE</code> the residuals are plotted versus the fitted values. |
| <code>whichx</code> | Indicates which predictor to plot against if there is more than one; i.e. indicates which column of the <code>x</code> matrix to use. If <code>vs.fit</code> is <code>TRUE</code> , <code>whichx</code> is ignored. |
| <code>shape</code> | The shape of the plotting symbol; possible values are "disc", "square", and "diamond". Partial matching is used so only the first few letters need be given. |
| <code>ngon</code> | The "disc" shape is actually a regular polygon; <code>ngon</code> specifies how many sides it should have; ignored if <code>shape</code> is not equal to "disc". |
| <code>size</code> | A scale factor to change the absolute sizes of the plotting symbols; values larger than 1 make the symbols larger; values less than 1 make them smaller. |
| <code>xlab</code> | The x label for the plot; defaults to "x". |
| <code>...</code> | Additional arguments to be passed to the polygon function which actually draws the plotting symbols. |

Details

This function is a "method" for `plot`. The plot produced is visually assessed by ignoring or discounting small symbols.

Side Effects

A residual plot is produced in whatever device is currently open.

ACKNOWLEDGEMENT

The idea of creating residual plots for regression mixtures by making the symbol size proportional to the associated probability is due to Adrian Baddeley of the University of Western Australia.

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Statist.* vol. 49, Part 3, pp. 371 – 384.

See Also

[bootcomp](#), [cband](#), [covmix](#), [mixreg](#), [plot.cband](#), [qq.mix](#), [resid.mix](#)

Examples

```
#See mixreg for examples.
```

| | |
|--------|---|
| qq.mix | <i>Draw a normal quantile-quantile plot of the residuals from a fitted mixture of linear regressions.</i> |
|--------|---|

Description

Draws a normal quantile-quantile plot using symbols whose size is proportional to the probability that the associated observation was generated by the associated component of the model.

Usage

```
qq.mix(object, xlim=NULL, ylim=NULL, shape="disc", ngon=20,
        size=1, ...)
```

Arguments

| | |
|--------|---|
| object | A list with components containing the residuals, the relevant probabilities, the predictors, and the observations, as returned by <code>resid.mix</code> ; probably the residuals should be standardized for use by <code>qq.mix</code> , i.e. <code>resid.mix</code> should be called with <code>std=TRUE</code> . |
| xlim | The limits (endpoints) of the x axis of the plot; chosen in the "usual" way by the plotting software if <code>xlim</code> is <code>NULL</code> . |
| ylim | The limits (endpoints) of the y axis of the plot; chosen in the "usual" way by the plotting software if <code>ylim</code> is <code>NULL</code> . |
| shape | The shape of the plotting symbol; possible values are "disc", "square", and "diamond". Partial matching is used so only the first few letters need be given. |
| ngon | The "disc" shape is actually a regular polygon; <code>ngon</code> specifies how many sides it should have; ignored if <code>shape</code> is not equal to "disc". |

`size` A scale factor to change the absolute sizes of the plotting symbols; values larger than 1 make the symbols larger than the default; values less than 1 make them smaller.

`...` Additional arguments to be passed to the polygon function which actually draws the plotting symbols.

Details

The plot produced is visually assessed by ignoring or discounting small symbols.

Side Effects

A normal quantile-quantile plot is drawn in whatever device is currently open.

ACKNOWLEDGEMENT

The idea of creating quantile-quantile plots for regression mixtures by making the symbol size proportional to the associated probability is due to Adrian Baddeley of the University of Western Australia

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Statist.* vol. 49, Part 3, pp. 371 – 384.

See Also

[bootcomp](#), [cband](#), [covmix](#), [mixreg](#), [plot.cband](#), [plot.mresid](#), [resid.mix](#)

Examples

```
#See mixreg for examples.
```

```
resid.mix
```

Calculate the residuals of a mixture of linear regressions.

Description

Calculates the residuals from each component of the mixture and the matrix of probabilities that each observation was generated by each component.

Usage

```
resid.mix(object, x, y, std=FALSE)
```

Arguments

| | |
|--------|---|
| object | List describing the fitted mixture of regressions as returned by mixreg. |
| x | The matrix of predictors to which the mixture has been fitted. If there is only one predictor then x may be a vector. |
| y | The vector of observed values to which the mixture has been fitted. |
| std | Logical argument; if TRUE then the residuals are standardized (by dividing them by their estimated standard deviation). |

Details

The calculation of the estimated standard deviations of the residuals is a little bit complicated since each component of the model is fitted using weighted regression in a setting in which the weights are NOT the reciprocals of error variances. See the reference below for more detail.

Value

A list (of class "mresid") with components

| | |
|-------|---|
| resid | The residuals of the model in a matrix. The k-th column of this matrix is the vector of residuals from the k-th component of the model. |
| gamma | The matrix of probabilities that each observation was generated by each component |
| x | The x argument of the call to resid.mix |
| y | The y argument of the call to resid.mix |

References

Turner, T. R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. Appl. Statist. vol. 49, Part 3, pp. 371 – 384.

See Also

[bootcomp](#), [cband](#), [covmix](#), [mixreg](#), [plot.cband](#), [plot.mresid](#), [qq.mix](#)

Examples

```
#See mixreg for examples.
```

Index

*Topic **datasets**

aphids, 1

*Topic **hplot**

plot.mresid, 10

qq.mix, 11

*Topic **models**

bootcomp, 2

cband, 4

covmix, 6

mixreg, 7

plot.cband, 9

plot.mresid, 10

qq.mix, 11

resid.mix, 13

*Topic **regression**

bootcomp, 2

cband, 4

covmix, 6

mixreg, 7

plot.cband, 9

plot.mresid, 10

qq.mix, 11

resid.mix, 13

aphids, 1

bootcomp, 2, 6–8, 10–13

cband, 4, 4, 7, 8, 10–13

covmix, 4, 6, 6, 8, 10–13

mixreg, 4, 6, 7, 7, 10–13

plot.cband, 4, 6–8, 9, 11–13

plot.mresid, 4, 6–8, 10, 10, 12, 13

qq.mix, 4, 6–8, 10, 11, 11, 13

resid.mix, 4, 6–8, 10–12, 13