# Package 'GB2group'

February 21, 2019

**Type** Package

**Title** Estimation of the Generalised Beta Distribution of the Second Kind from Grouped Data

**Version** 0.2.0

**Author** Vanesa Jorda <jordav@unican.es>, Jose Maria Sarabia <jose.sarabia@unican.es>, Markus Jäntti <markus.jantti@sofi.su.se>.

**Maintainer** Vanesa Jorda <jordav@unican.es>

**Depends** R (>= 3.1.0)

**Imports** GB2, minpack.lm, ineq, numDeriv

**Description** Estimation of the generalized beta distribution of the second kind (GB2) and related models using grouped data in form of income shares. The GB2 family is a general class of distributions that provides an accurate fit to income data. 'GB2group' includes functions to estimate the GB2, the Singh-Maddala, the Dagum, the Beta 2, the Lognormal and the Fisk distributions. 'GB2group' deploys two different econometric strategies to estimate these parametric distributions, non-linear least squares (NLS) and the generalised method of moments (GMM). Asymptotic standard errors are reported for the GMM estimates. Standard errors of the NLS estimates are obtained by Monte Carlo simulation. See Jorda et al. (2018) <arXiv:1808.09831> for a detailed description of the estimation procedure.

**License** GPL (>= 2)

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 5.0.1

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2019-02-21 13:20:03 UTC

# R topics documented:



---

`fit.plot` *Plot of the estimated Lorenz curve and the empirical income shares*

---

## Description

The function `fit.plot` plots the parametric Lorenz curve and the observed income shares used for the estimation of the income distributions belonging to the GB2 family.

## Usage

```
fit.plot(fit, fit.type = 1, fit.legend = FALSE, l.size = 0.7)
```

## Arguments

| | |
|---|---|
| `fit` | A character string "name" naming the object that contains the estimation of the parametric model for which the Lorenz curve is plotted. |
| `fit.type` | specifies the method used to estimate the parametric model. By default, `fit.type = 1`, which represents the Lorenz curve estimated by NLS. If `fit.type = 2`, the Lorenz curve belongs to the GMM estimation. |
| `fit.legend` | If `TRUE`, the graph includes a legend indicating the model for which the Lorenz curve is plotted. |
| `l.size` | determines the size of the legend. |

## Details

The function `fit.plot` represents the parametric Lorenz curves of some models of the GB2 family. Closed expressions of the Lorenz curves of these models are provided by Jorda et al. (2018). The parametric model must be estimated before representing the theoretical Lorenz curve. To do so, create an object containing the result of the following functions: `fitgroup.gb2`, `fitgroup.b2`, `fitgroup.da`, `fitgroup.sm`, `fitgroup.ln` or `fitgroup.f`. The name of this object is used (with quotations marks) as the first argument of `fit.plot` (see examples below). This function returns a plot with the Lorenz curve of the model estimated by NLS or GMM. More than one fit can be plotted, even when different sets of data are used. The legend indicates the distribution for which the Lorenz curve is represented.

## Value

the function `fit.plot` returns a graph with the theoretical Lorenz curves of the Generalised Beta of the Second Kind (GB2) family of income distributions and the income shares used for the estimation of these models.

## References

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

## Examples

```
fit.ln <- fitgroup.ln(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
fit.b2 <- fitgroup.b2(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
fit.plot(c("fit.ln", "fit.b2"), fit.legend = TRUE, l.size = 0.8)
```

---

# `fitgroup.b2` *Estimation of the Beta 2 distribution from group data*

---

## Description

The function `fitgroup.b2` implements the estimation of the Beta 2 distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.b2(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| `y` | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| `x` | Vector of population shares associated with the income shares provided by `y`. The default is a vector of equally sized population shares of the same length of `y`. |
| `gini.e` | specifies the survey Gini index expressed as a decimal. |
| `pc.inc` | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| `se.gmm` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is `FALSE`. |

| | |
|---|---|
| se.nls | If TRUE and the argument N is not NULL, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| se.scale | If TRUE and the argument N is not NULL, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| N | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| nrep | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| grid | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| rescale | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument rescale should be a positive real number which, by default, is set to 1000. |
| gini | if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM. |

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The Beta 2 distribution is a particular case of this model with $a = 1$, defined in terms of the cumulative distribution function as follows:

$$F(x; b, p, q) = B\left(\frac{x/b}{1 + x/b}; p; q\right); x > 0,$$

where $b$ is the scale parameter and $p, q$ are the shape parameters that define the heaviness of the tail and the skewness of the distribution.

The function fitgroup.da estimates the parameters of the Beta 2 distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, including The World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via nlsLM Conventionally, moment estimates of a restricted model are taken as initial values. A potential limitation of this method is that, as the dimensionality of the parameter space increases, it is more difficult to achieve global convergence. Although it seems quite intuitive that the moment estimates of the restricted model might be a good starting point, the optimization could converge to a local minimum, which might lead to inaccurate estimates of the parameters.

To provide different non-arbitrary combinations of starting values, we propose to define a sequence of numbers (provided by grid). For each value in this sequence, the moment estimate of one of the parameters is obtained using the survey Gini index, assuming that the other one is equal to the grid value. Using this procedure, we end up with as many combinations of initial values as values in the grid, which are used to obtain different sets of estimates, keeping the one with the smallest residual

sum of squares. Although we cannot ensure that our estimates belong to the global minimum, this procedure covers a larger proportion of the parameter space than just using the moment estimates of a particular sub-model. See Jorda et al. (2018) for details.

This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by `pc.inc`, to the population mean of the Beta 2 distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

`fitgroup.b2` also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of the parameters. The numerical optimisation is performed using `optim` with the BFGS method. If `optim` reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the Beta 2 distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the Beta 2 distribution is computed using the function `simgini.b2` which makes use of `gini.b2`. If this function reports NaN, the Gini index is estimated by Monte Carlo simulation of 10^6 samples of size N = 10^6.

## Value

the function `fitgroup.b2` returns the following objects:

- `nls.estimation` Matrix containing the parameters of the Beta 2 distribution estimated by NLS and, if `se.nls = TRUE`, their standard errors.
- `nls.rss` Residual sum of squares of the NLS estimation.
- `gmm.estimation` Matrix containing the parameters of the Beta 2 distribution estimated by GMM and, if `se.gmm = TRUE`, their standard errors.
- `gmm.rss` Weighted residual sum of squares of the GMM estimation.
- `gini.estimation` Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.b2(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

---

`fitgroup.da` *Estimation of the Dagum distribution from group data*

---

## Description

The function `fitgroup.da` implements the estimation of the Dagum distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.da(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| `y` | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| `x` | Vector of population shares associated with the income shares provided by `y`. The default is a vector of equally sized population shares of the same length of `y`. |
| `gini.e` | specifies the survey Gini index expressed as a decimal. |
| `pc.inc` | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| `se.gmm` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is `FALSE`. |

| | |
|---|---|
| se.nls | If TRUE and the argument N is not NULL, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| se.scale | If TRUE and the argument N is not NULL, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| N | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| nrep | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| grid | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| rescale | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument rescale should be a positive real number which, by default, is set to 1000. |
| gini | if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM. |

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The Dagum distribution is a particular case of this model with $q = 1$, defined in terms of the cumulative distribution function as follows:

$$F(x; a, b, p) = \left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p}$$

where $b$ is the scale parameter and $a, p$ are the shape parameters that define the heaviness of the tail and the skewness of the distribution.

The function fitgroup.da estimates the parameters of the Dagum distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, includingThe World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via nlsLM Conventionally, moment estimates of a restricted model are taken as initial values. A potential limitation of this method is that, as the dimensionality of the parameter space increases, it is more difficult to achieve global convergence. Although it seems quite intuitive that the moment estimates of the restricted model might be a good starting point, the optimization could converge to a local minimum, which might lead to inaccurate estimates of the parameters.

To provide different non-arbitrary combinations of starting values, we propose to define a sequence of numbers (provided by grid). For each value in this sequence, the moment estimate of one of the parameters is obtained using the survey Gini index, assuming that the other one is equal to the grid value. Using this procedure, we end up with as many combinations of initial values as values in the grid, which are used to obtain different sets of estimates, keeping the one with the smallest residual

sum of squares. Although we cannot ensure that our estimates belong to the global minimum, this procedure covers a larger proportion of the parameter space than just using the moment estimates of a particular sub-model. See Jorda et al. (2018) for details.

This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by `pc.inc`, to the population mean of the Dagum distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

`fitgroup.da` also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of the parameters. The numerical optimisation is performed using `optim` with the BFGS method. If `optim` reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the Dagum distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the Dagum distribution is computed using the function `simgini.da` which makes use of `gini.d`. If this function reports NaN, the Gini index is estimated by Monte Carlo simulation of 10^6 samples of size N = 10^6.

## Value

the function `fitgroup.da` returns the following objects:

- `nls.estimation` Matrix containing the parameters of the Dagum distribution estimated by NLS and, if `se.nls = TRUE`, their standard errors.
- `nls.rss` Residual sum of squares of the NLS estimation.
- `gmm.estimation` Matrix containing the parameters of the Dagum distribution estimated by GMM and, if `se.gmm = TRUE`, their standard errors.
- `gmm.rss` Weighted residual sum of squares of the GMM estimation.
- `gini.estimation` Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.da(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

---

`fitgroup.f` *Estimation of the Fisk distribution from group data*

---

## Description

The function `fitgroup.f` implements the estimation of the Fisk distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.f(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| `y` | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| `x` | Vector of population shares associated with the income shares provided by `y`. The default is a vector of equally sized population shares of the same length of `y`. |
| `gini.e` | specifies the survey Gini index expressed as a decimal. |
| `pc.inc` | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| `se.gmm` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is `FALSE`. |

| | |
|---|---|
| se.nls | If TRUE and the argument N is not NULL, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| se.scale | If TRUE and the argument N is not NULL, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| N | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| nrep | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| grid | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| rescale | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument rescale should be a positive real number which, by default, is set to 1000. |
| gini | if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM. |

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The Fisk distribution is a particular case of this model with $p = q = 1$, defined in terms of the cumulative distribution function as follows:

$$F(x; a, b) = \left(1 - \left(\frac{x}{b}\right)^{a}\right)^{-1}$$

where $b$ is the scale parameter and $a$ is the shape parameter.

The function fitgroup.f estimates the parameters of the Fisk distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, including The World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via nlsLM. We use the moment estimate of the $a$ parameter, obtained by equating the sample Gini index specified by gini.e to the population Gini index, as initial value. This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by pc.inc, to the population mean of the Fisk distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

fitgroup.f also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of

the parameters. The numerical optimisation is performed using optim with the BFGS method. If optim reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the Fisk distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the Fisk distribution is computed using the function simgini.f.If this function reports a value greater than 1, the Gini index is estimated by Monte Carlo simulation of 10^6 samples of size N = 10^6.

## Value

the function fitgroup.f returns the following objects:

- nls.estimation Matrix containing the parameters of the Fisk distribution estimated by NLS and, if se.nls = TRUE, their standard errors.
- nls.rss Residual sum of squares of the NLS estimation.
- gmm.estimation Matrix containing the parameters of the Fisk distribution estimated by GMM and, if se.gmm = TRUE, their standard errors.
- gmm.rss Weighted residual sum of squares of the GMM estimation.
- gini.estimation Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.f(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

`fitgroup.gb2` *Estimation of the GB2 distribution from group data*

## Description

The function `fitgroup.gb2` implements the estimation of the GB2 distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.gb2(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| `y` | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| `x` | Vector of population shares associated with the income shares provided by `y`. The default is a vector of equally sized population shares of the same length of `y`. |
| `gini.e` | specifies the survey Gini index expressed as a decimal. |
| `pc.inc` | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| `se.gmm` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is `FALSE`. |
| `se.nls` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size `N`. By default, this argument is `FALSE`. |
| `se.scale` | If `TRUE` and the argument `N` is not `NULL`, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size `N`. By default, this argument is `FALSE`. |
| `N` | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| `nrep` | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| `grid` | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| `rescale` | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument `rescale` should be a positive real number which, by default, is set to 1000. |

`gini` if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM.

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The GB2 distribution can be defined in terms of the cumulative distribution function as follows:

$$F(x; a, b, p, q) = B(\frac{(x/b)^a}{1 + (x/b)^a}; p; q); x > 0,$$

where $b$ is the scale parameter and $a, p, q$ are the shape parameters that define the heaviness of the tail and the skewness of the distribution.

The function `fitgroup.gb2` estimates the parameters of the GB2 distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, including The World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via `nlsLM` Conventionally, moment estimates of a restricted model are taken as initial values. A potential limitation of this method is that, as the dimensionality of the parameter space increases, it is more difficult to achieve global convergence. Although it seems quite intuitive that the moment estimates of the restricted model might be a good starting point, the optimization could converge to a local minimum, which might lead to inaccurate estimates of the parameters.

To provide several non-arbitrary combinations of starting values, we propose to set one of the parameters equal to one so that, the GB2 distribution becomes one of its three particular cases: the Dagum, the Beta 2 and the Singh-Maddala distributions. Then, we define a sequence of numbers (provided by `grid`). For each value in this sequence, the moment estimate of one of the parameters is obtained using the survey Gini index, assuming that the other one is equal to the grid value. Using this procedure, we end up with three times as many combinations of initial values as values in the grid, which are used to obtain different sets of estimates, keeping the estimation with the smallest residual sum of squares. Although we cannot ensure that our estimates belong to the global minimum, this procedure covers a larger proportion of the parameter space than just using the moment estimates of a particular sub-model. See Jorda et al. (2018) for details.

This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by `pc.inc`, to the population mean of the GB2 distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

`fitgroup.gb2` also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of the parameters. The numerical optimisation is performed using `optim` with the BFGS method. If `optim` reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making

possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the GB2 distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the GB2 distribution is computed using the function `gini.gb2`. If this function reports NA or 1, the Gini index is estimated by Monte Carlo simulation of 10^6 samples of size N = 10^6.

## Value

the function `fitgroup.gb2` returns the following objects:

- `nls.estimation` Matrix containing the parameters of the GB2 distribution estimated by NLS and, if `se.nls = TRUE`, their standard errors.
- `nls.rss` Residual sum of squares of the NLS estimation.
- `gmm.estimation` Matrix containing the parameters of the GB2 distribution estimated by GMM and, if `se.gmm = TRUE`, their standard errors.
- `gmm.rss` Weighted residual sum of squares of the GMM estimation.
- `gini.estimation` Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.gb2(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

`fitgroup.ln` *Estimation of the Lognormal distribution from group data*

## Description

The function `fitgroup.ln` implements the estimation of the Lognormal distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.ln(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| `y` | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| `x` | Vector of population shares associated with the income shares provided by `y`. The default is a vector of equally sized population shares of the same length of `y`. |
| `gini.e` | specifies the survey Gini index expressed as a decimal. |
| `pc.inc` | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| `se.gmm` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is `FALSE`. |
| `se.nls` | If `TRUE` and the argument `N` is not `NULL`, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size `N`. By default, this argument is `FALSE`. |
| `se.scale` | If `TRUE` and the argument `N` is not `NULL`, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size `N`. By default, this argument is `FALSE`. |
| `N` | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| `nrep` | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| `grid` | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| `rescale` | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument `rescale` should be a positive real number which, by default, is set to 1000. |

`gini` if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM.

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The Lognormal distribution is a limit case of this model, defined in terms of the cumulative distribution function as follows:

$$F(x;\mu,\sigma) = \Phi\left(\frac{log(x)-\mu}{\sigma}\right)$$

where $\mu$ is the scale parameter and $\sigma$ is the shape parameter.

The function `fitgroup.ln` estimates the parameters of the Lognormal distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, including The World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via `nlsLM`. We use the moment estimate of the $a$ parameter, obtained by equating the sample Gini index specified by `gini.e` to the population Gini index, as initial value. This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by `pc.inc`, to the population mean of the Lognormal distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

`fitgroup.ln` also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of the parameters. The numerical optimisation is performed using `optim` with the BFGS method. If `optim` reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the Lognormal distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the Lognormal distribution is computed using the function `gini.ln`. `gini.ln`.

## Value

the function `fitgroup.ln` returns the following objects:

- `nls.estimation` Matrix containing the parameters of the Lognormal distribution estimated by NLS and, if `se.nls = TRUE`, their standard errors.
- `nls.rss` Residual sum of squares of the NLS estimation.

- gmm.estimation Matrix containing the parameters of the Lognormal distribution estimated by GMM and, if se.gmm = TRUE, their standard errors.
- gmm.rss Weighted residual sum of squares of the GMM estimation.
- gini.estimation Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.ln(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

---

`fitgroup.sm` *Estimation of the Singh-Maddala distribution from group data*

---

## Description

The function fitgroup.sm implements the estimation of the Singh-Maddala distribution from group data in form of income shares using the non-linear least squares (NLS) and the generalised method of moments (GMM) estimators.

## Usage

```
fitgroup.sm(y, x = rep(1/length(y), length(y)), gini.e, pc.inc = NULL,
  se.gmm = FALSE, se.nls = FALSE, se.scale = FALSE, N = NULL,
  nrep = 10^3, grid = 1:20, rescale = 1000, gini = FALSE)
```

## Arguments

| | |
|---|---|
| y | Vector of (non-cumulative) income shares expressed as decimals or percentage. At least four data points are required to estimate the parameters of the income distribution. |
| x | Vector of population shares associated with the income shares provided by y. The default is a vector of equally sized population shares of the same length of y. |
| gini.e | specifies the survey Gini index expressed as a decimal. |
| pc.inc | specifies an estimate of per capita income. If not provided, the weighting matrix cannot be computed, hence GMM estimates will not be reported. |
| se.gmm | If TRUE and the argument N is not NULL, the standard errors of the shape parameters of the GMM estimation are computed using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016).See Jorda et al. (2018) for details. By default, this argument is FALSE. |
| se.nls | If TRUE and the argument N is not NULL, the standard errors of the NLS parameters are obtained using Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| se.scale | If TRUE and the argument N is not NULL, the standard error of the scale parameter of the GMM estimation is obtained by Monte Carlo simulation of random samples of size N. By default, this argument is FALSE. |
| N | Specifies the size of the sample from which the grouped data was generated. This information is required to compute the standard errors. |
| nrep | Number of samples to be drawn in the Monte Carlo simulation of the standard error of the NLS parameters and the scale parameter of the GMM estimation. |
| grid | A sequence of positive real numbers to be used as initial values using the algorithm developed by Jorda et al. (2018). |
| rescale | Rescalation factor of per capita income. Reescalation might help to invert the weight matrix when the scale is too large or too small. The argument rescale should be a positive real number which, by default, is set to 1000. |
| gini | if TRUE, reports an estimate of the Gini index using NLS and, if possible, GMM. |

## Details

The Generalised Beta of the Second Kind (GB2) is a general class of distributions that is acknowledged to provide an accurate fit to income data (McDonald 1984; McDonald and Mantrala,1995). The Singh-Maddala distribution is a particular case of this model with $p = 1$, defined in terms of the cumulative distribution function as follows:

$$F(x; a, b, q) = \left(1 - \left(\frac{x}{b}\right)^{a}\right)^{-q}$$

where $b$ is the scale parameter and $a, q$ are the shape parameters that define the heaviness of the tail and the skewness of the distribution.

The function fitgroup.sm estimates the parameters of the Singh-Maddala distribution using grouped data in form of income shares. These data must have been generated by setting the proportion of

observations in each group before sampling, so that the population proportions are fixed, whereas income shares are random variables. Examples of this type of data can be found in the largest datasets of grouped data, including The World Income Inequality Database (UNU-WIDER, 2017), PovcalNet (World Bank, 2018) or the World Wealth and Income Database (Alvaredo et al., 2018).

For NLS, numerical optimisation is achieved using the Levenberg-Marquardt Algorithm via `nlsLM` Conventionally, moment estimates of a restricted model are taken as initial values. A potential limitation of this method is that, as the dimensionality of the parameter space increases, it is more difficult to achieve global convergence. Although it seems quite intuitive that the moment estimates of the restricted model might be a good starting point, the optimization could converge to a local minimum, which might lead to inaccurate estimates of the parameters.

To provide different non-arbitrary combinations of starting values, we propose to define a sequence of numbers (provided by `grid`). For each value in this sequence, the moment estimate of one of the parameters is obtained using the survey Gini index, assuming that the other one is equal to the grid value. Using this procedure, we end up with as many combinations of initial values as values in the grid, which are used to obtain different sets of estimates, keeping the one with the smallest residual sum of squares. Although we cannot ensure that our estimates belong to the global minimum, this procedure covers a larger proportion of the parameter space than just using the moment estimates of a particular sub-model. See Jorda et al. (2018) for details.

This method, however, does not provide an estimate for the scale parameter because the Lorenz curve is independent to scale. The scale parameter is estimated by equating the sample mean, specified by `pc.inc`, to the population mean of the Singh-Maddala distribution. Because NLS does not use the optimal covariance matrix of the moment conditions, the standard errors of the parameters are obtained by Monte Carlo simulation. Please be aware that the estimation of the standard errors might take a long time, especially if the sample size is large.

`fitgroup.sm` also implements a two-stage GMM estimator. In the first stage, NLS estimates are obtained as described above, which are used to compute a first stage estimator of the weighting matrix. The weighting matrix is used in the second stage to obtain optimally weighted estimates of the parameters. The numerical optimisation is performed using `optim` with the BFGS method. If `optim` reports an error, the L-BFGS method is used. NLS estimates are used as initial values for the optimisation algorithm. The GMM estimation incorporates the optimal weight matrix, thus making possible to derive the asymptotic standard errors of the parameters using results from Beach and Davison(1983) and Hajargasht and Griffiths (2016). As in the NLS estimation, the scale parameter is obtained by matching the population mean of the Singh-Maddala distribution to the sample mean. Hence, the standard error of the scale parameter is estimated by Monte Carlo simulation.

The Gini index of the Singh-Maddala distribution is computed using the function `simgini.sm` which makes use of `gini.sm`. If this function reports NaN, the Gini index is estimated by Monte Carlo simulation of 10^6 samples of size N = 10^6.

## Value

the function `fitgroup.sm` returns the following objects:

- `nls.estimation` Matrix containing the parameters of the Singh-Maddala distribution estimated by NLS and, if `se.nls = TRUE`, their standard errors.
- `nls.rss` Residual sum of squares of the NLS estimation.
- `gmm.estimation` Matrix containing the parameters of the Singh-Maddala distribution estimated by GMM and, if `se.gmm = TRUE`, their standard errors.

- `gmm.rss` Weighted residual sum of squares of the GMM estimation.
- `gini.estimation` Vector with the survey Gini index and the estimated Gini indices using NLS and GMM whenever possible.

## References

Alvaredo, F., A. Atkinson, T. Piketty, E. Saez, and G. Zucman. The World Wealth and Income Database. http://www.wid.world.

Beach, C.M. and R. Davidson (1983): Distribution-free statistical inference with Lorenz curves and income shares, *The Review of Economic Studies*, 50, 723 - 735.

Hajargasht, G. and W.E. Griffiths (2016): Inference for Lorenz Curves, Tech. Rep., The University of Melbourne.

Jorda, V., Sarabia, J.M., & Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.

McDonald, J.B. (1984): Some Generalized Functions for the Size Distribution of Income, *Econometrica*, 52, 647 - 665.

McDonald, J.B. and A. Mantrala (1995): The distribution of personal income: revisited, *Journal of Applied Econometrics*, 10, 201 - 204.

UNU-WIDER (2018). World Income Inequality Database (WIID3.4). https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

World Bank (2018). PovcalNet Data Base. Washington, DC: World Bank. http://iresearch.worldbank.org/PovcalNet/home.aspx.

## Examples

```
fitgroup.sm(y = c(9, 13, 17, 22, 39), gini.e = 0.29)
```

# Index