

Package ‘REndo’

April 10, 2017

Type Package

Title Fitting Linear Models with Endogenous Regressors using Latent Instrumental Variables

Version 1.2

Date 2017-03-28

Author Raluca Gui,
Markus Meierer,
Rene Algesheimer

Maintainer Raluca Gui <raluca.gui@business.uzh.ch>

Description Fits linear models with endogenous regressor using latent instrumental variable approaches. The methods included in the package are Lewbel's (1997) <doi:10.2307/2171884> higher moments approach as well as Lewbel's (2012) <doi:10.1080/07350015.2012.643126> heteroskedasticity approach, Park and Gupta's (2012) <doi:10.1287/mksc.1120.0718> joint estimation method that uses Gaussian copula and Kim and Frees's (2007) <doi:10.1007/s11336-007-9008-1> multilevel generalized method of moment approach that deals with endogeneity in a multilevel setting. These are statistical techniques to address the endogeneity problem where no external instrumental variables are needed.
This version:
- solves an error occurring when using the multilevelIV() function with two levels, random intercept.
- returns the AIC and BIC for copulaCorrection() (method 1) and latentIV() methods.
- residuals and fitted values can be saved by users for latentIV() and copulaCorrection() methods.
- improves the summary methods for copulaCorrection() and multilevelIV() functions.

Imports optimx, mvtnorm, AER, e1071, stats, corpcor, lme4, gmm, lmtest, plyr, sandwich, Matrix

Depends methods

License GPL-3

RoxygenNote 6.0.1

Collate 'checkAssumptions.R' 'cop_Cont1.R' 'cop_Cont2.R'
'cop_Discrete.R' 'cop_bootsSE.R' 'cop_logLik.R' 'cop_pStar.R'

```
'copulaCorrection.R' 'copulaCorrection_class.R'
'f_hetErrorsIV.R' 'f_multilevelGMM.R' 'hetErrorsIV.R'
'hetErrorsIV_class.R' 'higherMomentsIV.R' 'internalIV.R'
'latentIV.R' 'latentIV_class.R' 'latentIV_loglik.R'
'methods_copulaCorrection.R' 'methods_hetErrorsIV.R'
'methods_latentIV.R' 'methods_multilevelIV.R' 'multilevelIV.R'
'mixedGMM.R' 'multilevelIV_class.R' 'tScores.R'
```

NeedsCompilation no
Repository CRAN
Date/Publication 2017-04-10 14:33:46 UTC

R topics documented:

boots	2
copulaCorrection	3
copulaPStar	6
dataCopC1	6
dataCopC2	7
dataCopDis	7
dataHigherMoments	8
dataLatentIV	9
hetErrorsIV	9
higherMomentsIV	11
internalIV	14
latentIV	15
mixedGMM	17
multilevelIV	19
tScores	20

Index	22
--------------	-----------

boots	<i>Bootstrapping Standard Errors</i>
-------	--------------------------------------

Description

Performs bootstrapping to obtain the standard errors of the estimates of the model with one continuous endogenous regressor estimated via maximum likelihood using the [copulaCorrection](#) function.

Usage

```
boots(bot, y, X, P, param, intercept = NULL, data = NULL)
```

Arguments

bot	number of bootstrap replicates.
y	the vector or matrix containing the dependent variable.
X	the data frame or matrix containing the regressors of the model, both exogenous and endogeneous. The last column/s should contain the endogeneous variable/s.
P	the vector containing the continuous, non-normally distributed endogeneous variable.
param	initial values for the parameters to be optimized over. See copulaCorrection for more details.
intercept	an optional parameter. The model is estimated by default with intercept. If no intercept is desired or the regressors matrix X contains already a column of ones, intercept should be given the value "no".
data	optional data frame or matrix containing the variables of the model.

Details

The function could be used only when there is a single endogenous regressor and method one is selected in [copulaCorrection](#). of the copulaCorrection function is used for estimation.

Value

Returns the standard errors of the estimates of the model using the copula method 1 described in Park and Gupta (2012). See Details section of [copulaCorrection](#).

See Also

[copulaCorrection](#)

copulaCorrection	<i>Fitting Linear Models Endogeneous Regressors using Gaussian Copula</i>
------------------	---

Description

Fits linear models with continuous or discrete endogeneous regressors using Gaussian copulas, method presented in Park and Gupta (2012). This is a statistical technique to address the endogeneity problem, where no external instrumental variables are needed. The important assumption of the model is that the endogeneous variables should NOT be normally distributed.

Usage

```
copulaCorrection(y,X,P,param,type, method, intercept, data)
```

Arguments

y	the vector or matrix containing the dependent variable.
X	the data frame or matrix containing the regressors of the model, both <i>exogenous</i> and <i>endogeneous</i> . The last column/s should contain the endogenous variable/s.
P	the matrix/vector containing the endogenous variables.
param	the vector of initial values for the parameters of the model to be supplied to the optimization algorithm. The parameters to be estimated are $\theta = \{b, a, \rho, \sigma\}$, where b are the parameters of the exogenous variables, a is the parameter of the endogenous variable, ρ is the parameter for the correlation between the error and the endogenous regressor, while σ is the standard deviation of the structural error.
type	the type of the endogenous regressor/s. It can take two values, "continuous" or "discrete".
method	the method used for estimating the model. It can take two values, "1" or "2", where "1" is the ML approach described in Park and Gupta (2012), and "2" is the equivalent OLS approach described in the same paper. "1" can be applied when there is just a single, continuous endogenous variable. With one discrete or more than one continuous endogenous regressors, the second method is applied by default.
intercept	optional parameter. The model is estimated by default with intercept. If no intercept is desired or the regressors matrix X contains already a column of ones, intercept should be given the value "no".
data	data frame or matrix containing the variables of the model.

Details

The maximum likelihood estimation is performed by the "BFGS" algorithm. When there are two endogenous regressors, there is no need for initial parameters since the method applied is by default the augmented OLS, which can be specified by using method two - "method="2".

Value

Depending on the method and the type of the variables, it returns the optimal values of the parameters and their standard errors. When the method one is used, the standard errors returned are obtained bootstrapping over 10 samples. If more bootstrapping samples are desired, the standard errors can be obtained using the [boots](#) function from the same package. The following are being returned and can be saved:

coefficients	the estimated coefficients.
standard errors	the corresponding estimated coefficients standard errors.
fitted.values	the fitted values.
residuals	the estimated residuals.
logLik	the estimated log likelihood value in the case of method 1.
AIC	Akaike Information Criterion in the case of method 1.
BIC	Bayesian Information Criterion in the case of method 1.

Author(s)

The implementation of the model by Raluca Gui based on the paper of Park and Gupta (2012).

References

Park, S. and Gupta, S., (2012), 'Handling Endogeneous Regressors by Joint Estimation Using Copulas', Marketing Science, 31(4), 567-86.

See Also

[higherMomentsIV](#)

Examples

```
#load dataset dataCopC1, where P is endogenous, continuous and not normally distributed
```

```
data(dataCopC1)
```

```
y <- dataCopC1[,1]
```

```
X <- dataCopC1[,2:5]
```

```
P <- dataCopC1[,5]
```

```
## Not run:
```

```
c1 <- copulaCorrection(y, X, P, type = "continuous", method = "1", intercept=FALSE)
```

```
summary(c1)
```

```
## End(Not run)
```

```
# an alternative model can be obtained using "method = "2"".
```

```
c12 <- copulaCorrection(y, X, P, type = "continuous", method = "2", intercept=FALSE)
```

```
summary(c12)
```

```
# with 2 endogeneous regressors no initial parameters are needed, the default is the augmented OLS.
```

```
data(dataCopC2)
```

```
y <- dataCopC2[,1]
```

```
X <- dataCopC2[,2:6]
```

```
P <- dataCopC2[,5:6]
```

```
c2 <- copulaCorrection(y, X, P, type = "continuous", method="2", intercept=FALSE)
```

```
summary(c2)
```

```
# load dataset with 1 discrete endogeneous variable.
```

```
# having more than 1 discrete endogenous regressor is also possible
```

```
data(dataCopDis)
```

```
y <- dataCopDis[,1]
```

```
X <- dataCopDis[,2:5]
```

```
P <- dataCopDis[,5]
```

```
c3 <- copulaCorrection(y, X, P, type = "discrete", intercept=FALSE, data = dataCopDis)
```

```
summary(c3)
```

copulaPStar

Inverse-Normal Distribution of the Empirical Distribution Function

Description

Computes the empirical distribution function of a variable and the inverse-normal distribution of ECDF.

Usage

```
copulaPStar(P)
```

Arguments

P - the variable for which the inverse-normal distribution of its empirical distribution function is needed.

Value

Returns the inverse-normal distribution of the empirical distribution function of variable P.

See Also

[copulaCorrection](#)

dataCopC1

Simulated Dataset

Description

A dataset with two exogenous regressors, X1,X2, and one endogenous, continuous regressor, P, having a T-distribution with 3 degrees of freedom. An intercept and a dependent variable, y, are also included. The true parameter values for the coefficients are: $b_0 = 2$, $b_1 = 1.5$, $b_2 = -3$ and the coefficient of the endogenous regressor is set to $a_1 = -1$.

Usage

```
data("dataCopC1")
```

Format

A data frame with 2500 observations on the following 5 variables.

y a numeric vector representing the dependent variable.

I a numeric vector representing the intercept.

X1 a numeric vector, normally distributed and exogenous.

X2 a numeric vector, normally distributed and exogenous.

P a numeric vector, continuous and endogenous having T-distribution with 3 degrees of freedom.

dataCopC2*Simulated Dataset*

Description

A dataset with two exogenous, normally distributed regressors, X1 and X2, two endogenous, continuous regressors, P1 and P2, having a T-distribution with 3 and 5 degrees of freedom respectively, with a correlation of 0.25. The correlation between P1 and the error was set at 0.33, while between P2 and the error, at 0.15. The dataset contains an intercept and the dependent variable, y. The true parameter value for the model: $y = b_0 + b_1 * X1 + b_2 * X2 + a_1 * P1 + a_2 * P2 + \text{eps}$, are: $b_0 = 2$, $b_1 = 1.5$, $b_2 = -3$, $a_1 = -1$, $a_2 = 0.8$.

Usage

```
data("dataCopC2")
```

Format

A data frame with 2500 observations on the following 6 variables.

y a numeric vector representing the dependent variable.

I a numeric vector representing the intercept.

X1 a numeric vector, normally distributed and exogenous.

X2 a numeric vector, normally distributed and exogenous.

P1 a numeric vector, continuous and endogenous having T-distribution with 3 degrees of freedom.

P2 a numeric vector, continuous and endogenous having T-distribution with 5 degrees of freedom.

dataCopDis*Simulated Dataset*

Description

A dataset with an intercept, two exogenous regressors and one endogenous, discrete variable, used for exemplifying the use of [copulaCorrection](#) function. The true parameter values are: $b_0 = 2$, $b_1 = 1.5$, $b_2 = -3$, and the coefficient of the endogenous variable is set to $a_1 = -1$. The correlation between the endogenous regressor P and the error term is 0.33. P has a Poisson distribution with $\lambda = 5$.

Usage

```
data("dataCopDis")
```

Format

A data frame with 2500 observations on the following 5 variables.

y a numeric vector representing the dependent variable.

I a numeric vector representing the intercept.

X1 a numeric vector, normally distributed and exogenous.

X2 a numeric vector, normally distributed and exogenous.

P a numeric vector, discrete and endogenous.

dataHigherMoments	<i>Simulated Dataset</i>
-------------------	--------------------------

Description

A dataset enclosing a dependent variable, *y*, two exogenous regressors, *X1* and *X2* and one endogenous variable, *P*. The endogenous regressor has to have a non-normal distribution for identification. The model is:

$$y = b_0 + b_1 * X1 + b_2 * X2 + a_1 * P + \epsilon$$

True parameter values are $b_0 = 2$, $b_1 = 1.5$, $b_2 = -3$, $a_1 = -1$.

Usage

```
data("dataHigherMoments")
```

Format

A data frame with 2500 observations on the following 4 variables.

y a numeric vector representing the dependent variable.

X1 a numeric vector, normally distributed and exogenous.

X2 a numeric vector, normally distributed and exogenous.

P a numeric vector, representing an endogenous regressor.

See Also

[higherMomentsIV](#)

dataLatentIV

*Simulated Dataset***Description**

A dataset with one endogenous, discrete regressor used for exemplifying the use of the Latent Instrumental Variable function [latentIV](#).

Usage

```
data("dataLatentIV")
```

Format

A data frame with 2500 observations on the following 3 variables.

y a numeric vector representing the dependent variable.

P a numeric vector representing a discrete and endogenous regressor.

Z a numeric vector representing the discrete, latent IV used to build *P*.

Details

The dataset was modeled according to the following equations:

$$P = g_0 * Z + nu$$

$$y = b_0 + a_1 * P + epsilon$$

where $g_0 = 2$, $b_0 = 3$ and $a_1 = -1$.

See Also

[internalIV](#)

hetErrorsIV

Fitting Linear Models with Endogenous Regressors using Heteroskedastic Covariance Restrictions

Description

This function estimates the model parameters and associated standard errors for a linear regression model with one or more endogenous regressors. Identification is achieved through heteroscedastic covariance restrictions within the triangular system as proposed in Lewbel(2012). The function `hetErrorsIV` builds on the `lewbel` function from the `ivlewbel` package. Changes have been made only to the printing and the summary of the function, as well as the name.

Usage

```
hetErrorsIV(formula, data, clustervar = NULL, robust = TRUE)
```

Arguments

formula	an object of class formula.
data	the data frame containing these data. This argument is mandatory.
clustervar	a character value naming the cluster on which to adjust the standard errors and test statistics.
robust	if TRUE the function reports standard errors and test statistics that have been corrected for the presence heteroscedasticity using White's method.

Details

The formula follows a four-part specification. The following formula is an example: $y_2 \sim y_1 \mid x_1 + x_2 + x_3 \mid x_1 + x_2 \mid z_1$. Here, y_2 is the dependent variable and y_1 is the endogenous regressor. The code $x_1 + x_2 + x_3$ represents the exogenous regressors whereas the third part $x_1 + x_2$ specifies the exogenous heteroscedastic variables from which the instruments are derived. The final part z_1 is optional, allowing the user to include traditional instrumental variables. If both `robust=TRUE` and `clustervar=TRUE`, the function overrides the `robust` command and computes clustered standard errors and test statistics adjusted to account for clustering. The function also computes partial F-statistics that indicate potentially weak identification. In cases where there is more than one endogenous regressor the Angrist-Pischke (2009) method for multivariate first-stage F-statistics is employed.

Value

Returns an object of class `hetREndo`, with the following components:

coefficients	a coefficient matrix with columns containing the estimates, associated standard errors, test statistics and p-values..
call	the matched call.
obs	the number of observations.
jtest	J-test for overidentifying restrictions.
ftest	Partial F-test statistics for weak IV detection.

Author(s)

The implementation of the model formula by based on the paper of Lewbel (2012).

References

Lewbel, A. (2012). Using Heteroskedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models, *Journal of Business & Economic Statistics*, **30**(1), 67-80.

Angrist, J. and Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricists Companion, Princeton University Press.

Fernihough, A. (2014). **ivlewb** package: Uses heteroscedasticity to estimate mismeasured and endogenous regressor models.

See Also

[lewbel](#), [higherMomentsIV](#), [latentIV](#), [copulaCorrection](#), [ivreg](#).

Examples

```
data(dataHigherMoments)
dataHetIV <- as.data.frame(dataHigherMoments)
hIV <- hetErrorsIV(y ~ P | X1 + X2 | X1 + X2, data = dataHetIV)
summary(hIV)
```

higherMomentsIV	<i>Fitting Linear Models with Endogenous Regressors using Lewbel's Higher Moments Approach</i>
-----------------	--

Description

Fits linear models with one endogenous regressor using internal instruments built using the approach described in Lewbel A. (1997). This is a statistical technique to address the endogeneity problem where no external instrumental variables are needed. The implementation allows the incorporation of external instruments if available. An important assumption for identification is that the endogenous variable has a skewed distribution.

Usage

```
higherMomentsIV(y, X, P, G = NULL, IIV = c("g", "gp", "gy", "yp", "p2",
      "y2"), EIV = NULL, data = NULL)
```

Arguments

y	the vector or matrix containing the dependent variable.
X	the data frame or matrix containing the exogenous regressors of the model.
P	the endogenous variables of the model as columns of a matrix or dataframe.
G	the functional form of G. It can take four values, x2, x3, lnx or 1/x. The last two forms are conditional on the values of the exogenous variables: greater than 1 or different from 0 respectively.
IIV	stands for "internal instrumental variable". It can take six values: g, gp, gy, yp, p2 or y2. Tells the function which internal instruments to be constructed from the data. See "Details" for further explanations.
EIV	stands for "external instrumental variable". It is an optional argument that lets the user specify any external variable(s) to be used as instrument(s).
data	optional data frame or list containing the variables in the model.

Details

Consider the model below:

$$Y_t = \beta_0 + \gamma' X_t + \alpha P_t + \epsilon_t \quad (1)$$

$$P_t = Z_t + \nu_t \quad (2)$$

The observed data consist of Y_t , X_t and P_t , while Z_t , ϵ_t , and ν_t are unobserved. The endogeneity problem arises from the correlation of P_t with the structural error, ϵ_t , since $E(\epsilon\nu) \neq 0$. The requirement for the structural and measurement error is to have mean zero, but no restriction is imposed on their distribution.

Let \bar{S} be the sample mean of a variable S_t and $G_t = G(X_t)$ for any given function G that has finite third own and cross moments. Lewbel(1997) proves that the following instruments can be constructed and used with 2SLS to obtain consistent estimates:

$$q_{1t} = (G_t - \bar{G}) \quad (3a)$$

$$q_{2t} = (G_t - \bar{G})(P_t - \bar{P}) \quad (3b)$$

$$q_{3t} = (G_t - \bar{G})(Y_t - \bar{Y}) \quad (3c)$$

$$q_{4t} = (Y_t - \bar{Y})(P_t - \bar{P}) \quad (3d)$$

$$q_{5t} = (P_t - \bar{P})^2 \quad (3e)$$

$$q_{6t} = (Y_t - \bar{Y})^2 \quad (3f)$$

Instruments in equations 3e and 3f can be used only when the measurement and the structural errors are symmetrically distributed. Otherwise, the use of the instruments does not require any distributional assumptions for the errors. Given that the regressors $G(X) = X$ are included as instruments, $G(X)$ should not be linear in X in equation 3a.

Let small letter denote deviation from the sample mean: $s_i = S_i - \bar{S}$. Then, using as instruments the variables presented in equations 3 together with 1 and X_t , the two-stage-least-squares estimation will provide consistent estimates for the parameters in equation 1 under the assumptions exposed in Lewbel(1997).

Value

Returns an object of class `ivreg`, with the following components:

<code>coefficients</code>	parameters estimates.
<code>residuals</code>	a vector of residuals.
<code>fitted.values</code>	a vector of predicted means.
<code>n</code>	number of observations.
<code>df.residual</code>	residual degrees of freedom for the fitted model.
<code>cov.unscaled</code>	unscaled covariance matrix for coefficients.
<code>sigma</code>	residual standard error.
<code>call</code>	the original function call.
<code>formula</code>	the model formula.

terms	a list with elements "regressors" and "instruments" containing the terms objects for the respective components.
levels	levels of the categorical regressors.
contrasts	the contrasts used for categorical regressors.
x	a list with elements "regressors", "instruments", "projected", containing the model matrices from the respective components. "projected" is the matrix of regressors projected on the image of the instruments.

Author(s)

The implementation of the model formula by Raluca Gui based on the paper of Lewbel (1997).

References

Lewbel, A. (1997). Constructing Instruments for Regressions with Measurement Error when No Additional Data Are Available, with An Application to Patents and R&D. *Econometrica*, **65**(5), 1201-1213.

See Also

[internalIV](#), [ivreg](#), [latentIV](#)

Examples

```
#load data
data(dataHigherMoments)
y <- dataHigherMoments[,1]
X <- cbind(dataHigherMoments[,2],dataHigherMoments[,3])
colnames(X) <- c("x1","x2")
P <- dataHigherMoments[,4]

# call higherMomentsIV with internal instrument yp = (Y - mean(Y))(P - mean(P))
h <- higherMomentsIV(y,X,P, G = "x2", IIV = "yp")

# build an additional instrument p2 = (P - mean(P))^2 using the internalIV() function
eiv <- internalIV(y,X,P, G="x2", IIV = "p2")

# use the additional variable as external instrument in higherMomentsIV()
h1 <- higherMomentsIV(y,X,P,G = "x2",IIV = "yp", EIV=eiv)
summary(h1)

# get the robust standard errors using robust.se() function from package ivpack
# library(ivpack)
# sder <- robust.se(h1)
```

internalIV

*Constructs Internal Instrumental Variables From Data***Description**

The function can be used to construct additional instruments to be supplied to [higherMomentsIV](#) as additional instruments in the "EIV" argument.

Usage

```
internalIV(y, X, P, G = NULL, IIV = c("g", "gp", "gy", "yp", "p2", "y2"),
  data = NULL)
```

Arguments

y	the vector or matrix containing the dependent variable.
X	the data frame or matrix containing the exogenous regressors of the model.
P	the endogenous variables of the model as columns of a matrix or dataframe.
G	the functional form of G. It can take four values, x2, x3, lnx or 1/x. The last two forms are conditional on the values of the exogenous variables: greater than 0 or different from 0 respectively.
IIV	the internal instrumental variable to be constructed. It can take six values, "g", "gp", "gy", "yp", "p2" or "y2". See the "Details" section of higherMomentsIV for a description of the internal instruments.
data	optional data frame or list containing the variables in the model.

Value

Returns a vector/matrix constructed from the data which can be used as instrumental variable either in [higherMomentsIV](#) or in any other function/algorithm making use of instruments.

References

Lewbel, A. (1997). "Lewbel, A. (1997). 'Constructing Instruments for Regressions with Measurement Error when No Additional Data Are Available, with An Application to Patents and R&D'. *Econometrica*, 65(5), 1201-1213."

See Also

[higherMomentsIV](#)

Examples

```
# load data
data(dataHigherMoments)
y <- dataHigherMoments[,1]
X <- cbind(dataHigherMoments[,2],dataHigherMoments[,3])
colnames(X) <- c("X1","X2")
P <- dataHigherMoments[,4]
# build an instrument gp = (G - mean(G))(P - mean(P)) using the internalIV() function
# with G = "x3" meaning G(X) = X^3
eiv <- internalIV(y,X,P, G="x3", IIV = "gp")
```

latentIV

Fitting Linear Models with one Endogenous Regressor using Latent Instrumental Variables

Description

Fits linear models with one endogenous regressor and no additional explanatory variables using the latent instrumental variable approach presented in Ebbes,P., Wedel,M., B"ockenholt, U., and Steerneman, A. G. M. (2005). This is a statistical technique to address the endogeneity problem where no external instrumental variables are needed. The important assumption of the model is that the latent variables are discrete with at least two groups with different means and the structural error is normally distributed.

Usage

```
latentIV(formula, param = NULL, data)
```

Arguments

formula	an object of type 'formula': a symbolic description of the model to be fitted. Example $\text{var1} \sim \text{var2}$, where var1 is a vector containing the dependent variable, while var2 is a vector containing the endogenous variable. An intercept is included by default.
param	a vector of initial values for the parameters of the model to be supplied to the optimization algorithm. In any model there are eight parameters. The first parameter is the intercept, then the coefficient of the endogenous variable followed by the means of the two groups of the latent IV (they need to be different, otherwise model is not identified), then the next three parameters are for the variance-covariance matrix. The last parameter is the probability of being in group 1. When not provided, initial parameters values are set equal to the OLS coefficients, the two group means are set to be equal to $\text{mean}(P)$ and $\text{mean}(P) + \text{sd}(P)$, the variance-covariance matrix has all elements equal to 1 while probG1 is set to equal 0.5.
data	data frame or list containing the variables of the model.

Details

Let's consider the model:

$$Y_t = \beta_0 + \alpha P_t + \epsilon_t$$

$$P_t = \pi' Z_t + \nu_t$$

where $t = 1, \dots, T$ indexes either time or cross-sectional units, Y_t is the dependent variable, P_t is a $k \times 1$ continuous, endogenous regressor, ϵ_t is a structural error term with mean zero and $E(\epsilon^2) = \sigma_\epsilon^2$, α and β are model parameters. Z_t is a 1×1 vector of instruments, and ν_t is a random error with mean zero and $E(\nu^2) = \sigma_\nu^2$. The endogeneity problem arises from the correlation of P and ϵ_t through $E(\epsilon\nu) = \sigma_{\epsilon\nu}$.

latentIV considers Z_t' to be a latent, discrete, exogenous variable with an unknown number of groups m and π is a vector of group means. It is assumed that Z is independent of the error terms ϵ and ν and that it has at least two groups with different means. The structural and random errors are considered normally distributed with mean zero and variance-covariance matrix Σ :

$$\Sigma = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma_\nu^2 \end{pmatrix}$$

The identification of the model lies in the assumption of the non-normality of P_t , the discreteness of the unobserved instruments and the existence of at least two groups with different means.

The method has been programmed such that the latent variable has two groups. Ebbes et al.(2005) show in a Monte Carlo experiment that even if the true number of the categories of the instrument is larger than two, latentIV estimates are approximately consistent. Besides, overfitting in terms of the number of groups/categories reduces the degrees of freedom and leads to efficiency loss. When provided by the user, the initial parameter values for the two group means have to be different, otherwise the model is not identified. For a model with additional explanatory variables a Bayesian approach is needed, since in a frequentist approach identification issues appear. The optimization algorithm used is BFGS.

Value

Returns the optimal values of the parameters as computed by maximum likelihood using BFGS algorithm.

coefficients	the value of the parameters for the intercept and the endogenous regressor as computed with maximum likelihood.
fitted.values	the fitted values.
means	the value of the parameters for the means of the two categories/groups of the latent instrumental variable.
sigma	the variance-covariance matrix sigma, where on the main diagonal are the variances of the structural error and that of the endogenous regressor and the off-diagonal terms are equal to the covariance between the errors.
probG1	the probability of being in group one. Since the model assumes that the latent instrumental variable has two groups, 1-probG1 gives the probability of group 2.
value	the value of the log-likelihood function corresponding to the optimal parameters.
AIC	Akaike Information Criterion.

BIC	Bayesian Information Criterion.
convcode	an integer code, the same as the output returned by <code>optimx</code> . 0 indicates successful completion. A possible error code is 1 which indicates that the iteration limit <code>maxit</code> had been reached.
hessian	a symmetric matrix giving an estimate of the Hessian at the solution found.

Author(s)

The implementation of the model formula by Raluca Gui based on the paper of Ebbes et al. (2005).

References

Ebbes, P., Wedel, M., B"ockenholt, U., and Steerneman, A. G. M. (2005). 'Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income'. *Quantitative Marketing and Economics*, **3**:365–392.

Examples

```
# load data
data(dataLatentIV)
# function call without any initial parameter values
l <- latentIV(y ~ P, data = dataLatentIV)
summary(l)
# function call with initial parameter values given by the user
l1 <- latentIV(y ~ P, c(2.9,-0.85,0,0.1,1,1,1,0.5), data = dataLatentIV)
summary(l1)
```

mixedGMM

Multilevel GMM Estimation

Description

Estimates multilevel models (max. 3 levels) employing the GMM approach presented in Kim and Frees (2007). One of the important features is that, using the hierarchical structure of the data, no external instrumental variables are needed, unlike traditional instrumental variable techniques.

Usage

```
mixedGMM(formula, endoVar, data = NULL)
```

Arguments

formula	an object of type 'formula': a symbolic description of the model to be fitted.
endoVar	a matrix or data frame containing the variables assumed to be endogenous.
data	optional data frame or list containing the variables of the model.

Details

When all model variables are assumed exogenous the GMM estimator is the usual GLS estimator. While the GLS model assumes all explanatory variables are uncorrelated with the random intercepts and slopes in the model, fixed effects models allow for endogeneity of all effects but sweeps out the random components as well as the explanatory variables at the same levels. The more general estimator presented here allows for some of the explanatory variables to be endogenous and uses this information to build internal instrumental variables. The multilevel GMM estimator uses both the between and within variations of the exogenous variables, but only the within variation of the variables assumed endogenous. The mixed GMM estimator equals the random effects estimator when all variables are assumed exogenous and is equal to the fixed effects estimator when all variables are assumed endogenous. In between different GMM estimators are obtained for different sets of endogenous/exogenous variables.

Value

returns the estimated coefficients together with their standard errors and p-values. It also returns the variance-covariance matrix and the weight matrix used in estimation.

<code>coefficients</code>	the estimated coefficients.
<code>coefSdErr</code>	the standard errors of the estimated coefficients.
<code>vcovMat</code>	the variance-covariance matrix.
<code>weighMat</code>	the weight matrix used in estimation.
<code>formula</code>	the formula of the estimated model.

Author(s)

The implementation of the model formula by Raluca Gui based on the paper of Kim and Frees (2007).

References

Kim, Jee-Seon and Frees, Edward W. (2007). 'Multilevel Modeling with Correlated Effects'. *Psychometrika*, 72(4), 505-533.

Examples

```
## Not run:
data(tScores)
endoVars <- tScores[,5:7]
formula <-
TLI ~ GRADE_3 + RETAINED + SWITCHSC + S_FREELU + FEMALE + BLACK + HISPANIC +
OTHER + C_COHORT + T_EXPERI + CLASS_SI + P_MINORI +
(1+GRADE_3 | CID) + (1 | SID)
model1<- mixedGMM(formula, endoVars, data=tScores)
coef(model1)

## End(Not run)
```

Description

Estimates multilevel models (max. 3 levels) employing the GMM approach presented in Kim and Frees (2007). One of the important features is that, using the hierarchical structure of the data, no external instrumental variables are needed, unlike traditional instrumental variable techniques.

Usage

```
multilevelIV(formula, endoVar, data = NULL)
```

Arguments

formula	an object of type 'formula': a symbolic description of the model to be fitted.
endoVar	a matrix or data frame containing the variables assumed to be endogenous.
data	optional data frame or list containing the variables of the model.

Details

When all model variables are assumed exogenous the GMM estimator is the usual GLS estimator. While the GLS model assumes all explanatory variables are uncorrelated with the random intercepts and slopes in the model, fixed effects models allow for endogeneity of all effects but sweeps out the random components as well as the explanatory variables at the same levels. The more general estimator presented here allows for some of the explanatory variables to be endogenous and uses this information to build internal instrumental variables. The multilevel GMM estimator uses both the between and within variations of the exogenous variables, but only the within variation of the variables assumed endogenous. The mixed GMM estimator equals the random effects estimator when all variables are assumed exogenous and is equal to the fixed effects estimator when all variables are assumed endogenous. In between different GMM estimators are obtained for different sets of endogenous/exogenous variables.

Value

returns the estimated coefficients together with their standard errors and p-values. It also returns the variance-covariance matrix and the weight matrix used in estimation.

coefficients	the estimated coefficients.
coefSdErr	the standard errors of the estimated coefficients.
vcovMat	the variance-covariance matrix.
weighMat	the weight matrix used in estimation.
formula	the formula of the estimated model.

Author(s)

The implementation of the model formula by Raluca Gui based on the paper of Kim and Frees (2007).

References

Kim, Jee-Seon and Frees, Edward W. (2007). 'Multilevel Modeling with Correlated Effects'. *Psychometrika*, 72(4), 505-533.

See Also

[internalIV](#), [ivreg](#), [latentIV](#), [copulaCorrection](#)

Examples

```
## Not run:
data(tScores)
endoVars <- tScores[,RETAINED:S_FREELU, with=FALSE]
formula1 <-
TLI ~ GRADE_3 + RETAINED + SWITCHSC + S_FREELU + FEMALE + BLACK + HISPANIC +
OTHER + C_COHORT + T_EXPERI + CLASS_SI + P_MINORI +
(1+GRADE_3 | CID) + (1 | SID)
model1<- multilevelIV(formula1, endoVars, data=tScores)
coef(model1)

## End(Not run)
```

tScores

Test scores of 3054 test scores of 1174 students in 60 schools

Description

A dataset containing student achievement scores on a statewide mathematics test between 1994 and 2000 in Dallas, Texas. Due to privacy protection the data is a simulated dataset based on the variance-covariance matrix of the real dataset. The variables are as follows:

- Intecept - intercept.
- SID - school ID. Runs from 1 to 60.
- CID - student ID. Runs from 1 to 1174.
- TLI - the mathematics test score.
- GRADE_3 - grade level minus 3.
- RETAINED - whether a student was retained in the same grade.
- SWITCHSC - whetehr a student switched schools during the previous year.
- S_FREELU - the proportion of students at each school who were eligible for a free or reduced-price lunch program.
- FEMALE - dummy variable, 1 if felmale student, 0 if male student.

- BLACK - dummy variable indicating whether the student is African-American.
- HISPANIC - dummy variable indicating whether the student is Hispanic.
- OTHER - dummy variable equal to 1 if the student is neither Caucasian, nor African-American, nor Hispanic.
- C_COHORT - the cohort the student belongs to.
- T_EXPERI - represents the average years of experience of teachers in a given school.
- CLASS_SI - represents the average class size in a given school.
- P_MINORI - represents the percentage of minority students in a given school.

Usage

```
data(tScores)
```

Format

A data frame with 3054 rows and 16 variables.

Index

- *Topic **GMM**
 - mixedGMM, [17](#)
 - multilevelIV, [19](#)
- *Topic **copula**
 - copulaPStar, [6](#)
- *Topic **datasets**
 - dataCopC1, [6](#)
 - dataCopC2, [7](#)
 - dataCopDis, [7](#)
 - dataHigherMoments, [8](#)
 - dataLatentIV, [9](#)
 - tScores, [20](#)
- *Topic **endogeneity**
 - copulaPStar, [6](#)
 - internalIV, [14](#)
 - mixedGMM, [17](#)
 - multilevelIV, [19](#)
- *Topic **endogenous**
 - hetErrorsIV, [9](#)
 - higherMomentsIV, [11](#)
 - latentIV, [15](#)
- *Topic **heteroskedastic**
 - hetErrorsIV, [9](#)
- *Topic **instrumental**
 - mixedGMM, [17](#)
 - multilevelIV, [19](#)
- *Topic **instruments**
 - hetErrorsIV, [9](#)
 - higherMomentsIV, [11](#)
 - internalIV, [14](#)
 - latentIV, [15](#)
 - mixedGMM, [17](#)
 - multilevelIV, [19](#)
- *Topic **latent**
 - higherMomentsIV, [11](#)
 - latentIV, [15](#)
- *Topic **lewbel**
 - hetErrorsIV, [9](#)
 - internalIV, [14](#)
- boots, [2, 4](#)
- copulaCorrection, [2, 3, 3, 6, 7, 11, 20](#)
- copulaPStar, [6](#)
- dataCopC1, [6](#)
- dataCopC2, [7](#)
- dataCopDis, [7](#)
- dataHigherMoments, [8](#)
- dataLatentIV, [9](#)
- hetErrorsIV, [9](#)
- higherMomentsIV, [5, 8, 11, 11, 14](#)
- internalIV, [9, 13, 14, 20](#)
- ivreg, [11, 13, 20](#)
- latentIV, [9, 11, 13, 15, 20](#)
- lewbel, [11](#)
- mixedGMM, [17](#)
- multilevelIV, [19](#)
- tScores, [20](#)