Package 'REndo'

August 29, 2016

Type Package

Title Fitting Linear Models with Endogenous Regressors when No External Instruments are Available

Version 1.0

Date 2015-12-01

Author Raluca Gui,
Markus Meierer,
Rene Algesheimer

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

Description Fits linear models with endogenous regressors using Internal Instrumental Variables (IIV) methods. These are statistical techniques to correct for endogeneity when no strong, valid external instrumental variables are available. The first version of the package offers two methods, the Latent Instrumental Variables (Ebbes et al., 2005) and Lewbel's higher moments approach (Lewbel, 1997). In a second version of the package, two other methods will be added, joint estimation using copulas (Park and Gupta, 2012) and multilevel GMM (Kim and Frees, 2007).

Imports stats,optimx,mvtnorm,AER,e1071,utils,methods

License GPL-3

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2015-12-09 00:21:19

R topics documented:

`dataHMLewbel` *Simulated Dataset*

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:

$$y = b0 + b1 * X1 + b2 * X2 + a1 * P + epsilon$$

True parameter values are `b0 = 2, b1 = 1.5, b2 = -3, a1 = -1`.

Usage

```
data("dataHMLewbel")
```

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

`hmlewbel`

`dataLIV` *Simulated Dataset*

Description

A dataset with one endogenous, discrete regressor used for exemplifying the use of the Latent Instrumental Variable function `liv`.

Usage

```
data("dataLIV")
```

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:

$$P = g0 * Z + nu$$

$$y = b0 + a1 * P + epsilon$$

where `g0 = 2`, `b0 = 3` and `a1 = -1`.

See Also

`liv`

`hmlewbel` *Fitting Linear Models with Endogenous Regressors using Lewbel's Higher Moments Approach*

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

```
hmlewbel(y, X, P, G = c("x2", "x3", "lnx", "1/x"), 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:

`coefficients`	parameters estimates.
`residulas`	a vector of residuals.
`fitted.values`	a vector of predicted means.
`n`	number of observations.

df.residual	residual degrees of freedom for the fitted model.
cov.unscaled	unscaled covariance matrix for coefficients.
sigma	residual standard error.
call	the original function call.
formula	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 foe 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, liv

Examples

```
#load data
data(dataHMLewbel)
y <- dataHMLewbel$y
X <- cbind(dataHMLewbel$X1,dataHMLewbel$X2)
colnames(X) <- c("X1","X2")
P <- dataHMLewbel$P

# call hmlewbel with internal instrument yp = (Y - mean(Y))(P - mean(P))
hmlewbel(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 hmlewbel()
h <- hmlewbel(y,X,P,G = "x2",IIV = "yp", EIV=eiv)
h$coefficients

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

internalIV *Constructs Internal Instrumental Variables From Data*

Description

The function can be used to construct additional instruments to be supplied to hmlewbel as additional instruments in the "EIV" argument.

Usage

```
internalIV(y, X, P, G = c("x2", "x3", "lnx", "1/x"), 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 hmlewbel 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 whcih can be used as instrumental variable either in hmlewbel 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

hmlewbel

Examples

```
# load data
data(dataHMLewbel)
y <- dataHMLewbel$y
X <- cbind(dataHMLewbel$X1,dataHMLewbel$X2)
colnames(X) <- c("X1","X2")
P <- dataHMLewbel$P
# 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")
```

`liv` *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

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

Arguments

`formula`	an object of type 'formula': a symbolic description of the model to be fitted. Example `var1 ~ var2`, where `var1` is a vector containing the dependent variable, while `var2` is a vector containing the endogenous variable.
`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 paramameters values are set equal to the OLS coefficients, the two group means are set to be equal to `mean(P)` and `mean(P) + sd(P)`, the variance-covariance matrix has all elements equal to 1 while `probG1` is set to equal 0.5.
`data`	optional 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,..,T` indexes either time or cross-sectional units, Y_t is the dependent variable, P_t is a `k x 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 `l x 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 $E(\epsilon\nu) = \sigma_{\epsilon\nu}$.

LIV 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 experiement that even if the true number of the categories of the instrument is larger than two, LIV 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 additonal 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`	returns the value of the parameters for the intercept and the endogenous regressor as computed with maximum likelihood.
`means`	returns the value of the parameters for the means of the two categories/groups of the latent instrumental variable.
`sigma`	returns 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`	returns 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.
`convcode`	an integer code, the same as the output returned by `optimx`. 0 indicates successful completion. A possible error code is 1 which indicates that the iteration limit maxit 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(dataLIV)
y <- dataLIV$y
P <- dataLIV$P
# function call without any initial parameter values
l  <- liv(y ~ P)
summary(l)
# function call with initial parameter values given by the user
l1 <- liv(y ~ P, c(2.9,-0.85,0,0.1,1,1,1,0.5))
summary(l1)
```

---

liv-class *liv S4 Object*

Description

This class is used to store and further analyze the results of the liv function

Slots

`formula` returns an object of type 'formula' as used in the call of the function. Example `var1 ~ var2`.

`coefficients` model's coefficients

`seCoefficients` the standard errors of the coefficients

`groupMeans` the coefficients of the means of the two groups considered to build the latent IV.

`seMeans` the standard errors of the groups means coefficients.

`sigma` the coefficients of the variance -covariance matrix.

`probG1` the coefficient of the probability of group 1.

`seProbG1` the standard error of the coefcients of probability of group 1.

`initValues` the initial parameter values.

`value` the value of the log=likelihood function computed at the optimal parameter values.

`convCode` the converge code.

`hessian` the hessian matrix.

Examples

```
getSlots("liv")
```

Index