Package 'ivgets'

July 15, 2024

Title General to Specific Modeling and Indicator Saturation in 2SLS Models

Version 0.1.2

Description Provides facilities of general to specific model selection for exogenous regressors in 2SLS models. Furthermore, indicator saturation methods can be used to detect outliers and structural breaks in the sample.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Depends R (>= 2.10), gets (>= 0.38), ivreg

Imports stats, stringr

Suggests covr, Formula, knitr, rmarkdown, testthat (>= 3.0.0)

URL https://github.com/jkurle/ivgets

BugReports https://github.com/jkurle/ivgets/issues

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author Kurle Jonas [aut, cre] (<https://orcid.org/0000-0003-2197-2012>)

Maintainer Kurle Jonas <mail@jonaskurle.com>

Repository CRAN

Date/Publication 2024-07-15 10:00:02 UTC

Contents

`artificial2sls` *Artificial data set for illustration.*

Description

A data set containing dependent variable, endogenous and exogenous regressors, and excluded instruments for 2SLS models. The structural error is also stored even though not observed in practice.

Usage

```
artificial2sls
```

Format

A data frame with 100 observations (rows) and 16 variables (columns):

name	variable description
y	dependent variable
x1	intercept
x2	relevant exogenous regressor
x3	irrelevant exogenous regressor
x4	irrelevant exogenous regressor
x5	irrelevant exogenous regressor
x6	irrelevant exogenous regressor
x7	irrelevant exogenous regressor
x8	irrelevant exogenous regressor
x9	irrelevant exogenous regressor
x10	irrelevant exogenous regressor
x11	relevant endogenous regressor
u	structural error (in practice unobserved)
z11	excluded instrument
z12	excluded instrument
id	unique observation identifier

`artificial2sls_contaminated` *Artificial data set with outliers for illustration.*

Description

A data set containing dependent variable, endogenous and exogenous regressors, and excluded instruments for 2SLS models. The structural error is also stored even though not observed in practice. Some errors are contaminated, making these observations outliers.

Usage

```
artificial2sls_contaminated
```

Format

A data frame with 100 observations (rows) and 16 variables (columns):

name	variable description
y	dependent variable
x1	intercept
x2	relevant exogenous regressor
x3	irrelevant exogenous regressor
x4	irrelevant exogenous regressor
x5	irrelevant exogenous regressor
x6	irrelevant exogenous regressor
x7	irrelevant exogenous regressor
x8	irrelevant exogenous regressor
x9	irrelevant exogenous regressor
x10	irrelevant exogenous regressor
x11	relevant endogenous regressor
u	structural error (in practice unobserved)
z11	excluded instrument
z12	excluded instrument
id	unique observation identifier

Details

The data frame has two additional attributes that store the indices of the outliers, `"outliers"`, and their magnitudes `"magnitude"`.

artificial2sls_shiny *Artificial data set without outliers prepared for shiny application.*

Description

Artificial data set without outliers prepared for shiny application.

Usage

```
artificial2sls_shiny
```

Format

A data frame with 100 observations (rows) and 17 variables (columns):

name	variable description
y	dependent variable
x1	intercept
x2	relevant exogenous regressor
x3	irrelevant exogenous regressor
x4	irrelevant exogenous regressor
x5	irrelevant exogenous regressor
x6	irrelevant exogenous regressor
x7	irrelevant exogenous regressor
x8	irrelevant exogenous regressor
x9	irrelevant exogenous regressor
x10	irrelevant exogenous regressor
x11	relevant endogenous regressor
u	structural error (in practice unobserved)
z11	excluded instrument
z12	excluded instrument
id	unique observation identifier
is.outlier	factor variable whether the observation is an outlier (`1`) or not (`0`)

extract_variables *Extract the first and second stage regressors of ivreg formula*

Description

`extract_variables` takes a formula object for `ivreg::ivreg()`, i.e. in a format of `y ~ x1 + x2 | x1 + z2` and extracts the different elements in a list.

Usage

```
extract_variables(formula)
```

Arguments

formula	A formula for the ivreg::ivreg function, i.e. in format y ~ x1 + x2 | z1 + z2.

Value

extract_variables returns a list with three components: $yvar stores the name of the dependent variable, $first the names of the regressors of the first stage and $second the names of the second stage regressors.

---

factory_indicators *Function factory for creating indicators from their names*

Description

factory_indicators creates a function that takes the name of an indicator and returns the corresponding indicator to be used in a regression. For user-specified indicators, it extracts the corresponding column from the uis matrix.

Usage

```
factory_indicators(n)
```

Arguments

n	An integer specifying the length of the indicators.

Details

Argument n should equal the number of observations in the data set which will be augmented with the indicators.

The created function takes a name of an indicator and the original uis argument that was used in indicator saturation and returns the indicator.

Value

factory_indicators returns a function called creator().

gets.ivreg *Gets modeling on an ivreg object*

Description

gets.ivreg conducts general-to-specific model selection on an ivreg object returned by ivreg::ivreg().

Usage

```
## S3 method for class 'ivreg'
gets(
  x,
  gum.result = NULL,
  t.pval = 0.05,
  wald.pval = t.pval,
  do.pet = TRUE,
  ar.LjungB = NULL,
  arch.LjungB = NULL,
  normality.JarqueB = NULL,
  include.gum = FALSE,
  include.1cut = FALSE,
  include.empty = FALSE,
  max.paths = NULL,
  turbo = FALSE,
  tol = 1e-07,
  max.regs = NULL,
  print.searchinfo = TRUE,
  alarm = FALSE,
  keep_exog = NULL,
  overid = NULL,
  weak = NULL,
  ...
)
```

Arguments

x	An object of class "ivreg", as returned by ivreg::ivreg().
gum.result	a list with the estimation results of the General Unrestricted Model (GUM), or NULL (default). If the estimation results of the GUM are already available, then re-estimation of the GUM is skipped if the estimation results are provided via this argument
t.pval	numeric value between 0 and 1. The significance level used for the two-sided regressor significance t-tests
wald.pval	numeric value between 0 and 1. The significance level used for the Parsimonious Encompassing Tests (PETs)

do.pet — logical. If TRUE (default), then a Parsimonious Encompassing Test (PET) against the GUM is undertaken at each regressor removal for the joint significance of all the deleted regressors along the current path. If FALSE, then a PET is not undertaken at each regressor removal

ar.LjungB — a two element vector or NULL (default). In the former case, the first element contains the AR-order, the second element the significance level. If NULL, then a test for autocorrelation is not conducted

arch.LjungB — a two element vector or NULL (default). In the former case, the first element contains the ARCH-order, the second element the significance level. If NULL, then a test for ARCH is not conducted

normality.JarqueB — NULL or a numeric value between 0 and 1. In the latter case, a test for non-normality is conducted using a significance level equal to normality.JarqueB. If NULL, then no test for non-normality is conducted

include.gum — logical. If TRUE, then the GUM (i.e. the starting model) is included among the terminal models. If FALSE (default), then the GUM is not included

include.1cut — logical. If TRUE, then the 1-cut model is added to the list of terminal models. If FALSE (default), then the 1-cut is not added, unless it is a terminal model in one of the paths

include.empty — logical. If TRUE, then the empty model is added to the list of terminal models. If FALSE (default), then the empty model is not added, unless it is a terminal model in one of the paths

max.paths — NULL (default) or an integer equal to or greater than 0. If NULL, then there is no limit to the number of paths. If an integer, for example 1, then this integer constitutes the maximum number of paths searched

turbo — logical. If TRUE, then (parts of) paths are not searched twice (or more) unnecessarily, thus yielding a significant potential for speed-gain. However, the checking of whether the search has arrived at a point it has already been comes with a slight computational overhead. Accordingly, if turbo=TRUE, then the total search time might in fact be higher than if turbo=FALSE. This happens if estimation is very fast, say, less than quarter of a second. Hence the default is FALSE

tol — numeric value (default = 1e-07). The tolerance for detecting linear dependencies in the columns of the variance-covariance matrix when computing the Wald-statistic used in the Parsimonious Encompassing Tests (PETs), see the qr.solve function

max.regs — integer. The maximum number of regressions along a deletion path. Do not alter unless you know what you are doing!

print.searchinfo — logical. If TRUE (default), then a print is returned whenever simiplification along a new path is started

alarm — logical. If TRUE, then a sound or beep is emitted (in order to alert the user) when the model selection ends

keep_exog — A numeric vector of indices or a character vector of names corresponding to the exogenous regressors in the data that should not be selected over. Default NULL

means that selection is over all exogenous regressors. If an intercept has been specified in the `formula` but is not already included in the `data`, then it can be kept by either including the index `0` or the character `"Intercept"`, respectively, as an element in `keep_exog`.

`overid` `NULL` if no Sargan test of overidentifying restrictions should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.

`weak` `NULL` if no weak instrument F-test on the first stage should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.

`...` Further arguments passed to or from other methods.

Value

Returns a list of class `"ivgets"` with three named elements. `$selection` stores the selection results from `getsFun` (including paths, terminal models, and best specification). `$final` stores the `ivreg` model object of the best specification or `NULL` if the GUM does not pass all diagnostics. `$keep` stores the names of the regressors that were not selected over, including the endogenous regressors, which are always kept.

`isat.ivreg` *Indicator saturation modeling on an ivreg object*

Description

`isat.ivreg` conducts indicator saturation model selection on an ivreg object returned by `ivreg::ivreg()`.

Usage

```
## S3 method for class 'ivreg'
isat(
  y,
  iis = TRUE,
  sis = FALSE,
  tis = FALSE,
  uis = FALSE,
  blocks = NULL,
  ratio.threshold = 0.8,
  max.block.size = 30,
  t.pval = 1/NROW(data),
  wald.pval = t.pval,
  do.pet = FALSE,
  ar.LjungB = NULL,
  arch.LjungB = NULL,
  normality.JarqueB = NULL,
  info.method = c("sc", "aic", "hq"),
```

```
  include.1cut = FALSE,
  include.empty = FALSE,
  max.paths = NULL,
  parallel.options = NULL,
  turbo = FALSE,
  tol = 1e-07,
  max.regs = NULL,
  print.searchinfo = TRUE,
  plot = NULL,
  alarm = FALSE,
  overid = NULL,
  weak = NULL,
  fast = FALSE,
  ...
)
```

Arguments

y	An object of class "ivreg", as returned by ivreg::ivreg().
iis	logical. If TRUE, impulse indicator saturation is performed.
sis	logical. If TRUE, step indicator saturation is performed.
tis	logical. If TRUE, trend indicator saturation is performed.
uis	a matrix of regressors, or a list of matrices. If a list, the matrices must have named columns that should not overlap with column names of any other matrices in the list.
blocks	NULL (default), an integer (the number of blocks) or a user-specified list that indicates how blocks should be put together. If NULL, then the number of blocks is determined automatically
ratio.threshold	Minimum ratio of variables in each block to total observations to determine the block size, default=0.8. Only relevant if blocks = NULL
max.block.size	Maximum size of block of variables to be selected over, default=30. Block size used is the maximum of given by either the ratio.threshold and max.block.size
t.pval	numeric value between 0 and 1. The significance level used for the two-sided regressor significance t-tests
wald.pval	numeric value between 0 and 1. The significance level used for the Parsimonious Encompassing Tests (PETs)
do.pet	logical. If TRUE, then a Parsimonious Encompassing Test (PET) against the GUM is undertaken at each regressor removal for the joint significance of all the deleted regressors along the current path. If FALSE (default), then a PET is not undertaken at each regressor removal. By default, the numeric value is the same as that of t.pval
ar.LjungB	a two-item list with names lag and pval, or NULL (default). In the former case lag contains the order of the Ljung and Box (1979) test for serial correlation in the standardised residuals, and pval contains the significance level. If

lag=NULL (default), then the order used is that of the estimated 'arx' object. If ar.Ljungb=NULL, then the standardised residuals are not checked for serial correlation

arch.LjungB — a two-item list with names lag and pval, or NULL (default). In the former case, lag contains the order of the Ljung and Box (1979) test for serial correlation in the squared standardised residuals, and pval contains the significance level. If lag=NULL (default), then the order used is that of the estimated 'arx' object. If arch.Ljungb=NULL, then the standardised residuals are not checked for ARCH

normality.JarqueB — NULL (the default) or a value between 0 and 1. In the latter case, a test for non-normality is conducted using a significance level equal to normality.JarqueB. If NULL, then no test for non-normality is conducted

info.method — character string, "sc" (default), "aic" or "hq", which determines the information criterion to be used when selecting among terminal models. The abbreviations are short for the Schwarz or Bayesian information criterion (sc), the Akaike information criterion (aic) and the Hannan-Quinn (hq) information criterion

include.1cut — logical. If TRUE, then the 1-cut model is included among the terminal models, if it passes the diagnostic tests, even if it is not equal to one of the terminals. If FALSE (default), then the 1-cut model is not included (unless it is one of the terminals)

include.empty — logical. If TRUE, then an empty model is included among the terminal models, if it passes the diagnostic tests, even if it is not equal to one of the terminals. If FALSE (default), then the empty model is not included (unless it is one of the terminals)

max.paths — NULL (default) or an integer indicating the maximum number of paths to search

parallel.options — NULL or an integer, i.e. the number of cores/threads to be used for parallel computing (implemented w/makeCluster and parLapply)

turbo — logical. If TRUE, then (parts of) paths are not searched twice (or more) unnecessarily, thus yielding a significant potential for speed-gain. However, the checking of whether the search has arrived at a point it has already been comes with a slight computational overhead. Accordingly, if turbo=TRUE, then the total search time might in fact be higher than if turbo=FALSE. This happens if estimation is very fast, say, less than quarter of a second. Hence the default is FALSE

tol — numeric value (default = 1e-07). The tolerance for detecting linear dependencies in the columns of the regressors (see qr function). Only used if LAPACK is FALSE (default)

max.regs — integer. The maximum number of regressions along a deletion path. It is not recommended that this is altered

print.searchinfo — logical. If TRUE (default), then a print is returned whenever simiplification along a new path is started, and whenever regressors are dropped due to exact multicolinearity

plot — NULL or logical. If TRUE, then the fitted values and the residuals of the final model are plotted after model selection. If NULL (default), then the value set by options determines whether a plot is produced or not.

alarm	logical. If TRUE, then a sound is emitted (in order to alert the user) when the model selection ends
overid	NULL if no Sargan test of overidentifying restrictions should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.
weak	NULL if no weak instrument F-test on the first stage should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.
fast	A logical value indicating whether to speed up the 2SLS estimation but providing less details. Requires overid == NULL and weak == NULL.
...	Further arguments passed to or from other methods.

Value

Returns a list of class "ivisat" with two named elements. $selection stores the selection results from isat (including paths, terminal models, and best specification). $final stores the ivreg model object of the best specification or NULL if the GUM does not pass all diagnostics.

---

ivDiag *User diagnostics for getsFun() and isat()*

Description

ivDiag provides several diagnostic tests for 2SLS models that can be used during model selection. Currently, a weak instrument F-test of the first stage(s) and the Sargan test of overidentifying restrictions on the validity of the instruments are implemented.

Usage

```
ivDiag(x, weak = FALSE, overid = FALSE)
```

Arguments

x	A list containing the estimation results of the 2SLS model. Must contain an entry $diag that contains the diagnostics provided by the ivreg::ivreg() command.
weak	A logical value whether to conduct weak instrument tests.
overid	A logical value whether to conduct the Sargan test of overidentifying restrictions.

Details

The resulting matrix also has an attribute named "is.reject.bad", which is a logical vector of length *m*. Each entry records whether a rejection of the test means that the diagnostics have failed or vice versa. The first entry refers to the first row, the second entry to the second row etc. However, this attribute is not used in the following estimations. Instead, the decision rule is specified inside the user.fun argument of gets::diagnostics(), which allows for a named entry $is.reject.bad.

Value

Returns a matrix with three columns named "statistic", "df", and "p-value" and *m* rows. Each row records these results for one of the tests, so the number of rows varies by the arguments specified and the model (e.g. how many first stages equations there are).

---

ivgets *General-to-specific modeling for 2SLS models*

Description

General-to-specific modeling for 2SLS models

Usage

```
ivgets(
  formula,
  data,
  gum.result = NULL,
  t.pval = 0.05,
  wald.pval = t.pval,
  do.pet = TRUE,
  ar.LjungB = NULL,
  arch.LjungB = NULL,
  normality.JarqueB = NULL,
  include.gum = FALSE,
  include.1cut = FALSE,
  include.empty = FALSE,
  max.paths = NULL,
  turbo = FALSE,
  tol = 1e-07,
  max.regs = NULL,
  print.searchinfo = TRUE,
  alarm = FALSE,
  keep_exog = NULL,
  overid = NULL,
  weak = NULL
)
```

Arguments

formula	A formula in the format y ~ x1 + x2 | z1 + z2.
data	A data frame with all necessary variables y, x, and z.
gum.result	a list with the estimation results of the General Unrestricted Model (GUM), or NULL (default). If the estimation results of the GUM are already available, then re-estimation of the GUM is skipped if the estimation results are provided via this argument

`t.pval` numeric value between 0 and 1. The significance level used for the two-sided regressor significance t-tests

`wald.pval` numeric value between 0 and 1. The significance level used for the Parsimonious Encompassing Tests (PETs)

`do.pet` logical. If TRUE (default), then a Parsimonious Encompassing Test (PET) against the GUM is undertaken at each regressor removal for the joint significance of all the deleted regressors along the current path. If FALSE, then a PET is not undertaken at each regressor removal

`ar.LjungB` a two element vector or NULL (default). In the former case, the first element contains the AR-order, the second element the significance level. If NULL, then a test for autocorrelation is not conducted

`arch.LjungB` a two element vector or NULL (default). In the former case, the first element contains the ARCH-order, the second element the significance level. If NULL, then a test for ARCH is not conducted

`normality.JarqueB`
NULL or a numeric value between 0 and 1. In the latter case, a test for non-normality is conducted using a significance level equal to `normality.JarqueB`. If NULL, then no test for non-normality is conducted

`include.gum` logical. If TRUE, then the GUM (i.e. the starting model) is included among the terminal models. If FALSE (default), then the GUM is not included

`include.1cut` logical. If TRUE, then the 1-cut model is added to the list of terminal models. If FALSE (default), then the 1-cut is not added, unless it is a terminal model in one of the paths

`include.empty` logical. If TRUE, then the empty model is added to the list of terminal models. If FALSE (default), then the empty model is not added, unless it is a terminal model in one of the paths

`max.paths` NULL (default) or an integer equal to or greater than 0. If NULL, then there is no limit to the number of paths. If an integer, for example 1, then this integer constitutes the maximum number of paths searched

`turbo` logical. If TRUE, then (parts of) paths are not searched twice (or more) unnecessarily, thus yielding a significant potential for speed-gain. However, the checking of whether the search has arrived at a point it has already been comes with a slight computational overhead. Accordingly, if `turbo=TRUE`, then the total search time might in fact be higher than if `turbo=FALSE`. This happens if estimation is very fast, say, less than quarter of a second. Hence the default is FALSE

`tol` numeric value (`default = 1e-07`). The tolerance for detecting linear dependencies in the columns of the variance-covariance matrix when computing the Wald-statistic used in the Parsimonious Encompassing Tests (PETs), see the `qr.solve` function

`max.regs` integer. The maximum number of regressions along a deletion path. Do not alter unless you know what you are doing!

`print.searchinfo`
logical. If TRUE (default), then a print is returned whenever simiplification along a new path is started

alarm	logical. If TRUE, then a sound or beep is emitted (in order to alert the user) when the model selection ends
keep_exog	A numeric vector of indices or a character vector of names corresponding to the exogenous regressors in the data that should not be selected over. Default NULL means that selection is over all exogenous regressors. If an intercept has been specified in the formula but is not already included in the data, then it can be kept by either including the index 0 or the character "Intercept", respectively, as an element in keep_exog.
overid	NULL if no Sargan test of overidentifying restrictions should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.
weak	NULL if no weak instrument F-test on the first stage should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.

Value

Returns a list of class "ivgets" with three named elements. $selection stores the selection results from getsFun (including paths, terminal models, and best specification). $final stores the ivreg model object of the best specification or NULL if the GUM does not pass all diagnostics. $keep stores the names of the regressors that were not selected over, including the endogenous regressors, which are always kept.

---

ivisat *Indicator saturation modeling for 2SLS models*

Description

Indicator saturation modeling for 2SLS models

Usage

```
ivisat(
  formula,
  data,
  iis = TRUE,
  sis = FALSE,
  tis = FALSE,
  uis = FALSE,
  blocks = NULL,
  ratio.threshold = 0.8,
  max.block.size = 30,
  t.pval = 1/NROW(data),
  wald.pval = t.pval,
  do.pet = FALSE,
  ar.LjungB = NULL,
```

```
  arch.LjungB = NULL,
  normality.JarqueB = NULL,
  info.method = c("sc", "aic", "hq"),
  include.1cut = FALSE,
  include.empty = FALSE,
  max.paths = NULL,
  parallel.options = NULL,
  turbo = FALSE,
  tol = 1e-07,
  max.regs = NULL,
  print.searchinfo = TRUE,
  plot = NULL,
  alarm = FALSE,
  overid = NULL,
  weak = NULL,
  fast = FALSE
)
```

Arguments

formula	A formula in the format y ~ x1 + x2 | z1 + z2.
data	A data frame with all necessary variables y, x, and z.
iis	logical. If TRUE, impulse indicator saturation is performed.
sis	logical. If TRUE, step indicator saturation is performed.
tis	logical. If TRUE, trend indicator saturation is performed.
uis	a matrix of regressors, or a list of matrices. If a list, the matrices must have named columns that should not overlap with column names of any other matrices in the list.
blocks	NULL (default), an integer (the number of blocks) or a user-specified list that indicates how blocks should be put together. If NULL, then the number of blocks is determined automatically
ratio.threshold	Minimum ratio of variables in each block to total observations to determine the block size, default=0.8. Only relevant if blocks = NULL
max.block.size	Maximum size of block of variables to be selected over, default=30. Block size used is the maximum of given by either the ratio.threshold and max.block.size
t.pval	numeric value between 0 and 1. The significance level used for the two-sided regressor significance t-tests
wald.pval	numeric value between 0 and 1. The significance level used for the Parsimonious Encompassing Tests (PETs)
do.pet	logical. If TRUE, then a Parsimonious Encompassing Test (PET) against the GUM is undertaken at each regressor removal for the joint significance of all the deleted regressors along the current path. If FALSE (default), then a PET is not undertaken at each regressor removal. By default, the numeric value is the same as that of t.pval

`ar.LjungB` a two-item list with names `lag` and `pval`, or NULL (default). In the former case `lag` contains the order of the Ljung and Box (1979) test for serial correlation in the standardised residuals, and `pval` contains the significance level. If `lag=NULL` (default), then the order used is that of the estimated 'arx' object. If `ar.Ljungb=NULL`, then the standardised residuals are not checked for serial correlation

`arch.LjungB` a two-item list with names `lag` and `pval`, or NULL (default). In the former case, `lag` contains the order of the Ljung and Box (1979) test for serial correlation in the squared standardised residuals, and `pval` contains the significance level. If `lag=NULL` (default), then the order used is that of the estimated 'arx' object. If `arch.Ljungb=NULL`, then the standardised residuals are not checked for ARCH

`normality.JarqueB`
NULL (the default) or a value between 0 and 1. In the latter case, a test for non-normality is conducted using a significance level equal to `normality.JarqueB`. If `NULL`, then no test for non-normality is conducted

`info.method` character string, "sc" (default), "aic" or "hq", which determines the information criterion to be used when selecting among terminal models. The abbreviations are short for the Schwarz or Bayesian information criterion (sc), the Akaike information criterion (aic) and the Hannan-Quinn (hq) information criterion

`include.1cut` logical. If `TRUE`, then the 1-cut model is included among the terminal models, if it passes the diagnostic tests, even if it is not equal to one of the terminals. If FALSE (default), then the 1-cut model is not included (unless it is one of the terminals)

`include.empty` logical. If `TRUE`, then an empty model is included among the terminal models, if it passes the diagnostic tests, even if it is not equal to one of the terminals. If FALSE (default), then the empty model is not included (unless it is one of the terminals)

`max.paths` NULL (default) or an integer indicating the maximum number of paths to search

`parallel.options`
NULL or an integer, i.e. the number of cores/threads to be used for parallel computing (implemented w/`makeCluster` and `parLapply`)

`turbo` logical. If `TRUE`, then (parts of) paths are not searched twice (or more) unnecessarily, thus yielding a significant potential for speed-gain. However, the checking of whether the search has arrived at a point it has already been comes with a slight computational overhead. Accordingly, if `turbo=TRUE`, then the total search time might in fact be higher than if `turbo=FALSE`. This happens if estimation is very fast, say, less than quarter of a second. Hence the default is `FALSE`

`tol` numeric value (default = 1e-07). The tolerance for detecting linear dependencies in the columns of the regressors (see `qr` function). Only used if LAPACK is FALSE (default)

`max.regs` integer. The maximum number of regressions along a deletion path. It is not recommended that this is altered

`print.searchinfo`
logical. If `TRUE` (default), then a print is returned whenever simiplification along a new path is started, and whenever regressors are dropped due to exact multicolinearity

plot	NULL or logical. If TRUE, then the fitted values and the residuals of the final model are plotted after model selection. If NULL (default), then the value set by options determines whether a plot is produced or not.
alarm	logical. If TRUE, then a sound is emitted (in order to alert the user) when the model selection ends
overid	NULL if no Sargan test of overidentifying restrictions should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.
weak	NULL if no weak instrument F-test on the first stage should be used as a diagnostic check for model selection or a numeric value between 0 and 1. In the latter case, the test is conducted using this value as the significance level.
fast	A logical value indicating whether to speed up the 2SLS estimation but providing less details. Requires overid == NULL and weak == NULL.

Value

Returns a list of class "ivisat" with two named elements. $selection stores the selection results from isat (including paths, terminal models, and best specification). $final stores the ivreg model object of the best specification or NULL if the GUM does not pass all diagnostics.

`ivregFun` *User estimator ivreg for getsFun() and isat()*

Description

ivregFun calls ivreg::ivreg() in a format that is suitable for the model selection function gets::getsFun() and for the indicator saturation function gets::isat().

Usage

```
ivregFun(y, x, z, formula, tests, fast = FALSE)
```

Arguments

y	A numeric vector with no missing values.
x	A matrix or NULL.
z	A numeric vector or matrix.
formula	A formula in the format y ~ x1 + x2 | z1 + z2.
tests	A logical value whether to calculate the ivreg::summary.ivreg() diagnostics.
fast	A logical value whether to speed up the 2SLS estimation but providing less details. Requires tests == FALSE.

Details

For the required outputs of user-specified estimators, see the article "User-Specified General-to-Specific and Indicator Saturation Methods" by Genaro Sucarrat, published in the R Journal: https://journal.r-project.org/archive/2021/RJ-2021-024/index.html

Value

A list with entries needed for model selection via `gets::getsFun()` or `gets::isat()`.

`new_formula` *Takes ivreg formula and returns formula compatible with model selection*

Description

`new_formula` takes a formula object for `ivreg::ivreg()`, i.e. in a format of `y ~ x1 + x2 | x1 + z2`, and returns a list with element suitable for model selection. For example, it updates the data by creating an intercept if specified in the formula, checks for collinearity among the regressors, and updates the formula accordingly.

Usage

```
new_formula(formula, data, keep_exog)
```

Arguments

`formula`	A formula for the ivreg::ivreg function, i.e. in format `y ~ x1 + x2 | z1 + z2`.
`data`	A data frame.
`keep_exog`	A numeric vector of indices or a character vector of names corresponding to the exogenous regressors in the `data` that should not be selected over. Default `NULL` means that selection is over all exogenous regressors. If an intercept has been specified in the `formula` but is not already included in the `data`, then it can be kept by either including the index `0` or the character `"Intercept"`, respectively, as an element in `keep_exog`.

Value

A list with several named elements. Component `$fml` stores the new baseline formula that will be used for model selection. Components `y`, `x`, and `z` store the data of the dependent variable, structural regressors, and excluded instruments. The entries `$depvar`, `$x1`, `$x2`, `$z1`, and `$z2` contain the names of the dependent variable, endogenous and exogenous regressors, included and excluded instruments. `$dx1`, `$dx2`, `$dz1`, `$dz2` store the dimensions of the respective variables. Finally, `$keep` and `$keep.names` contain the indices and names of the regressors that will not be selected over.

Index