Title

> **poivregress —** Partialing-out lasso instrumental-variables regression

Description	Quick start	Menu	Syntax
Options	Remarks and examples	Stored results	Methods and formulas
Reference	Also see		

Description

poivregress fits a lasso instrumental-variables linear regression model and reports coefficients along with standard errors, test statistics, and confidence intervals for specified covariates of interest. The covariates of interest may be endogenous or exogenous. The partialing-out method is used to estimate effects for these variables and to select from potential control variables and instruments to be included in the model.

Quick start

Estimate a coefficient for endogenous d1 in a linear regression of y on d1, and include x1 to x100 as potential control variables and z1 to z100 as potential instruments to be selected by lassos

```
poivregress y (d1 = z1-z100), controls(x1-x100)
```

Same as above, and estimate the coefficient for the exogenous d2

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100)
```

Use cross-validation (CV) instead of a plugin iterative formula to select the optimal λ^* in each lasso

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100) selection(cv)
```

Same as above, and set a random-number seed for reproducibility

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100) selection(cv) ///
        rseed(28)
```

Specify CV for the lasso for y only, with the stopping rule criterion turned off

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100)  ///
        lasso(y, selection(cv), stop(0))
```

Same as above, but apply the option to the lassos for y, d2, and d1

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100)  ///
        lasso(*, selection(cv), stop(0))
```

Compute lassos beyond the CV minimum to get full coefficient paths, knots, etc.

```
poivregress y d2 (d1 = z1-z100), controls(x1-x100)  ///
        lasso(*, selection(cv, alllambdas))
```

Menu

Statistics > Lasso > Lasso inferential models > Continuous outcomes > Partialing-out IV model

Syntax

> poivregress *depvar* [*exovars*] (*endovars* = *instrumvars*) [*if*] [*in*],
>
>   controls([(*alwaysvars*)] *othervars*) [*options*]

Coefficients and standard errors are estimated for the exogenous variables, *exovars*, and the endogenous variables, *endovars*. The set of instrumental variables, *instrumvars*, may be high dimensional.

options	Description
Model	
* controls([(*alwaysvars*)] *othervars*)	*alwaysvars* and *othervars* are control variables for *depvar*, *exovars*, and *endovars*; *instrumvars* are an additional set of control variables that apply only to the *endovars*; *alwaysvars* are always included; lassos choose whether to include or exclude *othervars*
selection(plugin)	use a plugin iterative formula to select an optimal value of the lasso penalty parameter λ^* for each lasso; the default
selection(cv)	use CV to select an optimal value of the lasso penalty parameter λ^* for each lasso
selection(adaptive)	use adaptive lasso to select an optimal value of the lasso penalty parameter λ^* for each lasso
selection(bic)	use BIC to select an optimal value of the lasso penalty parameter λ^* for each lasso
sqrtlasso	use square-root lassos
missingok	after fitting lassos, ignore missing values in any *instrumvars* or *othervars* not selected, and include these observations in the final model
SE/Robust	
vce(*vcetype*)	*vcetype* may be robust (the default) or cluster *clustvar*
Reporting	
level(#)	set confidence level; default is level(95)
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Optimization	
[no]log	display or suppress an iteration log
verbose	display a verbose iteration log
rseed(#)	set random-number seed
Advanced	
lasso(*varlist*, *lasso_options*)	specify options for the lassos for variables in *varlist*; may be repeated
sqrtlasso(*varlist*, *lasso_options*)	specify options for square-root lassos for variables in *varlist*; may be repeated
reestimate	refit the model after using lassoselect to select a different λ^*
noheader	do not display the header on the coefficient table
coeflegend	display legend instead of statistics

*controls() is required.

exovars, *endovars*, *instrumvars*, *alwaysvars*, and *othervars* may contain factor variables. Base levels of factor variables cannot be set for *instrumvars*, *alwaysvars*, and *othervars*. See [U] **11.4.3 Factor variables**.

collect is allowed; see [U] **11.1.10 Prefix commands**.

reestimate, noheader, and coeflegend do not appear in the dialog box.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

> Model

controls($\big[$(*alwaysvars*)$\big]$ *othervars*) specifies the set of control variables, which control for omitted variables. Control variables are also known as confounding variables. *alwaysvars* are variables that are always to be included in lassos. *alwaysvars* are optional. *othervars* are variables that lassos will choose to include or exclude. The instrumental variables, *instrumvars*, are an additional set of control variables, but they apply only to the *endovars*. controls() is required.

> poivregress fits lassos for *depvar* and each one of the *exovars* and *endovars*. The control variables for the lassos for *depvar* and *exovars* are *alwaysvars* (always included) and *othervars* (lasso will include or exclude). The control variables for lassos for *endovars* are *exovars* (always included), *alwaysvars* (always included), *instrumvars* (lasso will include or exclude), and *othervars* (lasso will include or exclude).

selection(plugin | cv | adaptive | bic) specifies the selection method for choosing an optimal value of the lasso penalty parameter λ^* for each lasso or square-root lasso estimation. Separate lassos are estimated for *depvar* and each variable in *varsofinterest*. Specifying selection() changes the selection method for all of these lassos. You can specify different selection methods for different lassos using the option lasso() or sqrtlasso(). When lasso() or sqrtlasso() is used to specify a different selection method for the lassos of some variables, they override the global setting made using selection() for the specified variables.

> selection(plugin) is the default. It selects λ^* based on a "plugin" iterative formula dependent on the data. See [LASSO] **lasso options**.

> selection(cv) selects the λ^* that gives the minimum of the CV function. See [LASSO] **lasso options**.

> selection(adaptive) selects λ^* using the adaptive lasso selection method. It cannot be specified when sqrtlasso is specified. See [LASSO] **lasso options**.

> selection(bic) selects the λ^* that gives the minimum of the BIC function. See [LASSO] **lasso options**.

sqrtlasso specifies that square-root lassos be done rather than regular lassos. The option lasso() can be used with sqrtlasso to specify that regular lasso be done for some variables, overriding the global sqrtlasso setting for these variables. See [LASSO] **lasso options**.

missingok specifies that, after fitting lassos, the estimation sample be redefined based on only the nonmissing observations of variables in the final model. In all cases, any observation with missing values for *depvar*, *exovars*, *endovars*, *instrumvars*, *alwaysvars*, and *othervars* is omitted from the estimation sample for the lassos. By default, the same sample is used for calculation of the coefficients of the *exovars* and *endovars* and their standard errors.

When `missingok` is specified, the initial estimation sample is the same as the default, but the sample used for the calculation of the coefficients of the *exovars* and *endovars* can be larger. Now observations with missing values for any *instrumvars* and *othervars* not selected will be added to the estimation sample (provided there are no missing values for any of the variables in the final model).

`missingok` may produce more efficient estimates when data are missing completely at random. It does, however, have the consequence that estimation samples can change when selected variables differ in models fit using different selection methods. That is, when *instrumvars* and *othervars* contain missing values, the estimation sample for a model fit using the default `selection(plugin)` will likely differ from the estimation sample for a model fit using, for example, `selection(cv)`.

⌐ SE/Robust ⌐

`vce(vcetype)` specifies the type of standard error reported, which includes types that are robust to some kinds of misspecification (`robust`) and that allow for intragroup correlation (`cluster` *clustvar*); see [R] **vce_option**.

When `vce(cluster clustvar)` is specified, all lassos also account for clustering. For each lasso, this affects how the log-likelihood function is computed and how the sample is split in cross-validation; see *Methods and formulas* in [LASSO] **lasso**. Specifying `vce(cluster clustvar)` may lead to different selected controls and therefore to different point estimates for your variable of interest when compared to the estimation that ignores clustering.

⌐ Reporting ⌐

`level(#)`; see [R] **Estimation options**.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(%fmt)`, `pformat(%fmt)`, `sformat(%fmt)`, and `nolstretch`; see [R] **Estimation options**.

⌐ Optimization ⌐

[`no`]`log` displays or suppresses a log showing the progress of the estimation. By default, one-line messages indicating when each lasso estimation begins are shown. Specify `verbose` to see a more detailed log.

`verbose` displays a verbose log showing the iterations of each lasso estimation. This option is useful when doing `selection(cv)` or `selection(adaptive)`. It allows you to monitor the progress of the lasso estimations for these selection methods, which can be time consuming when there are many *othervars* specified in `controls()` or many *instrumvars*.

`rseed(#)` sets the random-number seed. This option can be used to reproduce results for `selection(cv)` and `selection(adaptive)`. The default selection method `selection(plugin)` does not use random numbers. `rseed(#)` is equivalent to typing `set seed #` prior to running `poivregress`. See [R] **set seed**.

⌐ Advanced ⌐

`lasso(varlist, lasso_options)` lets you set different options for different lassos, or advanced options for all lassos. You specify a *varlist* followed by the options you want to apply to the lassos for these variables, where *varlist* consists of one or more variables from *depvar*, *exovars*, or *endovars*. `_all` or `*` may be used to specify *depvar* and all *exovars* and *endovars*. This option is repeatable as long as different variables are given in each specification. *lasso_options*

are `selection(...)`, `grid(...)`, `stop(#)`, `tolerance(#)`, `dtolerance(#)`, and `cvtolerance(#)`. When `lasso(varlist, selection(...))` is specified, it overrides any global `selection()` option for the variables in *varlist*. It also overrides the global `sqrtlasso` option for these variables. See [LASSO] **lasso options**.

`sqrtlasso(varlist, lasso_options)` works like the option `lasso()`, except square-root lassos for the variables in *varlist* are done rather than regular lassos. *varlist* consists of one or more variables from *depvar*, *exovars*, or *endovars*. This option is repeatable as long as different variables are given in each specification. *lasso_options* are `selection(...)`, `grid(...)`, `stop(#)`, `tolerance(#)`, `dtolerance(#)`, and `cvtolerance(#)`. When `sqrtlasso(varlist, selection(...))` is specified, it overrides any global `selection()` option for the variables in *varlist*. See [LASSO] **lasso options**.

The following options are available with `poivregress` but are not shown in the dialog box:

`reestimate` is an advanced option that refits the `poivregress` model based on changes made to the underlying lassos using `lassoselect`. After running `poivregress`, you can select a different λ^* for one or more of the lassos estimated by `poivregress`. After selecting λ^*, you type `poivregress, reestimate` to refit the `poivregress` model based on the newly selected λ's.

`reestimate` may be combined only with reporting options.

`noheader` prevents the coefficient table header from being displayed.

`coeflegend`; see [R] **Estimation options**.

Remarks and examples

`poivregress` performs partialing-out lasso instrumental-variables linear regression. This command estimates coefficients, standard errors, and confidence intervals and performs tests for variables of interest, both exogenous and endogenous, while using lassos to select from among potential control variables and instruments.

The instrumental-variables linear regression model is

$$y = \mathbf{d}\boldsymbol{\alpha}'_d + \mathbf{f}\boldsymbol{\alpha}'_f + \mathbf{x}\boldsymbol{\beta}' + \epsilon$$

where $\mathbf{d}$ are the endogenous variables, $\mathbf{f}$ are the exogenous variables for which we wish to make inferences, and $\mathbf{x}$ are the potential control variables from which the lassos select. In addition, lassos select from potential instrumental variables, $\mathbf{z}$. `poivregress` reports estimated coefficients for $\boldsymbol{\alpha}_d$ and $\boldsymbol{\alpha}_f$. However, partialing-out does not provide estimates of the coefficients on the control variables or their standard errors. No estimation results can be reported for $\boldsymbol{\beta}$.

For an introduction to the partialing-out lasso method for inference, as well as the double-selection and cross-fit partialing-out methods, see [LASSO] **Lasso inference intro**.

Examples that demonstrate how to use `poivregress` and the other lasso inference commands are presented in [LASSO] **Inference examples**. In particular, we recommend reading *1 Overview* for an introduction to the examples and to the `vl` command, which provides tools for working with the large lists of variables that are often included when using lasso methods. See *2 Fitting and interpreting inferential models* for comparisons of the different methods of fitting inferential models that are available in Stata. See *6 Fitting an inferential model with endogenous covariates* for examples and discussion specific to models that account for endogenous covariates.

If you are interested in digging deeper into the lassos that are used to select controls, see *5 Exploring inferential model lassos* in [LASSO] **Inference examples**.

Stored results

poivregress stores the following in e():

Scalars
    e(N)   number of observations
    e(N_clust)   number of clusters
    e(k_varsofinterest) number of variables of interest
    e(k_controls)   number of potential control variables
    e(k_controls_sel)   number of selected control variables
    e(k_inst)   number of potential instruments
    e(k_inst_sel)   number of selected instruments
    e(df)   degrees of freedom for test of variables of interest
    e(chi2) χ^2
    e(p) p-value for test of variables of interest
    e(rank)   rank of e(V)

Macros
    e(cmd)   poivregress
    e(cmdline)   command as typed
    e(depvar)   name of dependent variable
    e(lasso_depvars)   names of dependent variables for all lassos
    e(varsofinterest)   variables of interest
    e(controls)   potential control variables
    e(controls_sel)   selected control variables
    e(exog)   exogenous variables
    e(endog)   endogenous variables
    e(inst)   potential instruments
    e(inst_sel)   selected instruments
    e(model)   linear
    e(title)   title in estimation output
    e(clustvar)   name of cluster variable
    e(chi2type)   Wald; type of χ^2 test
    e(vce)   *vcetype* specified in vce()
    e(vcetype)   title used to label Std. err.
    e(rngstate)   random-number state used
    e(properties)   b V
    e(predict)   program used to implement predict
    e(select_cmd)   program used to implement lassoselect
    e(marginsnotok)   predictions disallowed by margins
    e(asbalanced)   factor variables fvset as asbalanced
    e(asobserved)   factor variables fvset as asobserved

Matrices
    e(b)   coefficient vector
    e(V)   variance–covariance matrix of the estimators

Functions
    e(sample)   marks estimation sample

In addition to the above, the following is stored in r():

Matrices
    r(table)   matrix containing the coefficients with their standard errors, test statistics, p-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

poivregress implements partialing-out lasso instrumental-variables regression described in Chernozhukov, Hansen, and Spindler (2015). The model is

$$y = \mathbf{d}\boldsymbol{\alpha}'_d + \mathbf{f}\boldsymbol{\alpha}'_f + \mathbf{x}\boldsymbol{\beta}' + \epsilon$$

where $\mathbf{d}$ contains the J_d endogenous covariates of interest, $\mathbf{f}$ contains the J_f exogenous covariates of interest, and $\mathbf{x}$ contains the p_x controls. We also have p_z outside instrumental variables, denoted by $\mathbf{z}$, that are correlated with $\mathbf{d}$ but not with ϵ. The number of controls in $\mathbf{x}$ and the number of instruments in $\mathbf{z}$ can be large and, in theory, can grow with the sample size; however, the number of nonzero elements in $\boldsymbol{\beta}$ and nonzero coefficients of $\mathbf{z}$ must not be too large, which is to say that the model must be sparse. See *Stata commands for inference* in [LASSO] **Lasso intro** for a discussion on what it means for the model to be sparse.

In the following algorithm, each lasso can choose the lasso penalty parameter (λ^*) using the plugin estimator, adaptive lasso, or CV. The heteroskedastic plugin estimator for the linear lasso is the default.

Partialing-out lasso instrumental-variables regression algorithm

1. Perform a linear lasso of y on $\mathbf{x}$, and denote the selected controls by $\widetilde{\mathbf{x}}_y$.

   Fit a linear regression of y on $\widetilde{\mathbf{x}}_y$, and denote the residual for the ith observation from this regression by $\widetilde{\rho}_i$.

2. For $j = 1, \ldots, J_d$, perform a linear lasso of d_j on $\mathbf{f}$, $\mathbf{x}$, and $\mathbf{z}$, denote the selected controls by $\widetilde{\mathbf{x}}_{d,j}$, and denote the selected instruments by $\widetilde{\mathbf{z}}_j$.

   Fit a linear regression of d_j on $\mathbf{f}$, $\widetilde{\mathbf{x}}_{d,j}$, and $\widetilde{\mathbf{z}}_j$, and denote the linear prediction from this regression by $\widehat{d}_j$.

   Perform a linear lasso of $\widehat{d}_j$ on the controls $\mathbf{x}$, and denote the selected controls by $\check{\mathbf{x}}_{d,j}$.

   Fit a linear regression of $\widehat{d}_j$ on $\check{\mathbf{x}}_{d,j}$, let $\check{\boldsymbol{\beta}}_j$ be the estimated coefficients, and denote the residuals from this regression by $\check{d}_j$, with its ith observation denoted by $\check{d}_{j,i}$.

   Also compute the "residuals" for the levels

   $$\widetilde{d}_j = d_j - \check{\mathbf{x}}_j \check{\boldsymbol{\beta}}_j$$

   and denote its ith observation by $\widetilde{d}_{j,i}$.

3. For $j = 1, \ldots, J_f$, perform a linear lasso of f_j on the controls $\mathbf{x}$, and denote the selected controls by $\widetilde{\mathbf{x}}_{f,j}$.

   Fit a linear regression of f_j on $\widetilde{\mathbf{x}}_{f,j}$, and denote the residual for the ith observation by $\widetilde{f}_{j,i}$.

4. Form the vector of instruments

   $$\mathbf{w}_i = \left(\check{d}_{1,i}, \ldots, \check{d}_{J_d,i}, \widetilde{f}_{1,i}, \ldots, \widetilde{f}_{J_f,i} \right)$$

5. Form the vector of partialed-out covariates

   $$\mathbf{p}_i = \left(\widetilde{d}_{1,i}, \ldots, \widetilde{d}_{J_d,i}, \widetilde{f}_{1,i}, \ldots, \widetilde{f}_{J_f,i} \right)$$

6. Compute $\widehat{\alpha}$ by solving the following $J_d + J_f$ sample-moment equations.

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i'(\widetilde{\rho}_i - \mathbf{p}_i\widehat{\alpha}') = \mathbf{0}$$

7. The variance for $\widehat{\alpha}$ is estimated by

$$\widehat{\mathbf{Var}}(\widehat{\alpha}) = \frac{1}{n}\left(\widehat{\mathbf{J}}_0^{-1}\right)\widehat{\mathbf{\Psi}}\left(\widehat{\mathbf{J}}_0^{-1}\right)'$$

where

$$\widehat{\mathbf{J}}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i'\mathbf{p}_i$$

$$\widehat{\mathbf{\Psi}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i\widehat{\psi}_i'$$

$$\widehat{\psi}_i = \mathbf{w}_i'(\widetilde{\rho}_i - \mathbf{p}_i\widehat{\alpha}')$$

See *Methods and formulas* in [LASSO] **lasso** for details on how the lassos in steps 1, 2, and 3 choose their penalty parameters (λ^*).

Reference

Chernozhukov, V., C. B. Hansen, and M. Spindler. 2015. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics* 7: 649–688. https://doi.org/10.1146/annurev-economics-012315-015826.

Also see

[LASSO] **lasso inference postestimation** — Postestimation tools for lasso inferential models

[LASSO] **xpoivregress** — Cross-fit partialing-out lasso instrumental-variables regression

[R] **ivregress** — Single-equation instrumental-variables regression

[U] **20 Estimation and postestimation commands**