ivcrc – Correlated random coefficient instrumental variables regression
Author: David Benson

Syntax

ivcrc *depvar* [*varlist*$_1$] (*varname*$_2$ = *varlist*$_{iv}$) [if] [in] [, options]

Optional exogenous regressors are included in *varlist*$_1$. The basic endogenous regressor is *varname*$_2$, and the (excluded) instrumental variables are included in *varlist*$_{iv}$. For the multivariate module, *varname*$_2$ may be a list of basic endogenous variables as in the usual *ivregress* syntax.

Options	Description
dendog(*varlist*)	Specify derived endogenous variables.
bootstrap()	Bootstrap confidence intervals and standard errors; default is no standard errors. May specify typical bootstrap options in (), e.g. *reps(#)*.
normal	Report normal distribution confidence intervals, rather than empirical percentiles, for bootstrap standard errors.
integral(*numlist* [, report])	Specifies the grid space on $(0, 1)$ for numerical integration; default grid is $(.01, .02, ..., .99)$. The grid may be composed of multiple non-overlapping ascending subsets. Sub-option: specifying integral(*numlist*, report) returns the average coefficient for each subset as well as over the entire space. Sub-option: together with varcoef(), user specifies the support for kernel weights in a varying coefficients model.
kernel(*string*)	Choose alternative kernel functions; default is the uniform kernel. Other options: triangle, biweight, triweight, cosine, epanechnikov, or gaussian.
bandwidth(*numlist*)	Bandwidth of kernel; default is 0.05. If multiple values are specified, estimates for each bandwidth are reported. Sub-option: together with varcoef(), specifies the user's bandwidth for a varying coefficients model.
quantiles(*integer*)	Number (minus 1) of evenly spaced quantiles for computation of conditional rank statistic; default is 100.
generate(*varname* [, replace])	Save conditional rank statistic to *varname* in the working dataset; ignored when bootstrapping.
varcoef(*varlist*)	Allows for coefficients conditioned on covariates specified in *varlist*, an alternative to conditioning on the rank of basic endogenous variables. Options integral() and bandwidth() must be specified together with varcoef().

1 Notes

This program estimates random coefficient models of the form

$$Y = B_0 + \sum_{j=1}^{d_x} B_j X_j + \sum_{j=1}^{d_1} B_{d_x+j} Z_{1j}$$

as found in Masten and Torgovitsky (2016). In their notation, there are d_x endogenous variables X, d_1 included exogenous variables Z_1, and d_2 excluded exogenous variables Z_2. The module estimates the average of the coefficient vector B, notated β.

1.1 Computation of conditional rank statistic

The rank of the endogenous regressor $X := \mathit{varname}_2$ conditional on instruments $Z := (\ \mathit{varlist}_{iv}, \mathit{varlist}_1)$ is computed in two steps. First, $Q-1$ evenly spaced quantiles of the endogenous variable are selected and for each $q = \frac{1}{Q}, ..., \frac{Q-1}{Q}$ a linear quantile regression is estimated (see qreg command) of the form

$$X_i = Z_i'\theta_q + e_{qi}$$

The estimated coefficients $\hat{\theta}_q$ are used to predict $\hat{X}(q, Z_i)$ and indicator variables $1(\hat{X}(q, Z_i) \leq X_i)$, yielding $Q-1$ indicators for each observation. Second, each observation's rank statistic is computed as the row-mean over the indicators:

$$\hat{R}_i(X_i|Z_i) = \frac{1}{Q-1}\sum_q 1(\hat{X}(q, Z_i) \leq X_i)$$

Computation of the rank is the same whether there is a single basic endogenous regressor or multiple basic endogenous regressors.

1.2 Computation of $\hat{\beta}$: single basic endogenous variable

When there is a single basic endogenous regressor, the average β of the endogenous coefficient B is computed in two steps. First, a finite grid or set of grids $\mathcal{R}$ on $(0,1)$ is chosen, the default being $\mathcal{R} = \{.01, .02, ..., .99\}$. Given the estimated rank $\hat{R}_i$ and bandwidth parameter h, at each point $r \in \mathcal{R}$ a weighted least squares regression is run using kernel weights $K(\frac{\hat{R}_i - r}{h})$, producing a local coefficient $\hat{\beta}(r) = \mathbb{E}(B|R = r)$. Second, the estimates $\hat{\beta}(r)$ are numerically integrated over $\mathcal{R}$ to compute $\hat{\beta}$. If $\mathcal{R}$ is a single set, then the integral is approximated by

$$\hat{\beta}(\mathcal{R}) = \frac{1}{|\mathcal{R}|}\sum_{r \in \mathcal{R}} \hat{\beta}(r)$$

where $|\mathcal{R}|$ is the number of points on the grid. If $\mathcal{R}$ is a set of disjoint grids $\mathcal{R} = (\mathcal{R}_1, ..., \mathcal{R}_n)$, each $\mathcal{R}_j$ a subset of $(0,1)$ with lebesgue measure $\lambda(\mathcal{R}_j)$, then coefficients $\hat{\beta}(\mathcal{R}_j) = \mathbb{E}(B|R \in \mathcal{R}_j)$ are computed as above and the integral is approximated by

$$\hat{\beta}(\mathcal{R}) = \frac{1}{\lambda(\mathcal{R}_1) + ... + \lambda(\mathcal{R}_n)} \sum_{\mathcal{R}_j \in \mathcal{R}} \lambda(\mathcal{R}_j)\hat{\beta}(\mathcal{R}_j)$$

1.3 Computation of $\hat{\beta}$: multiple basic endogenous variables

When there is more than one basic endogenous variable, the module does not use a grid method to compute $\hat{\beta}$. If the d_x basic endogenous variables are correlated, then the support of the rank statistics will be a strict subset of $[0,1]^{d_x}$. Moreover, if d_x is large and/or the desired grid for averaging is very fine, then there may be fewer observations in the data than there are points at which to estimate each $\hat{\beta}(r)$. For these reasons the multivariate module takes $\mathcal{R}$ to be the empirical support of $\hat{R}_i$, and computes $\beta(r)$ at each empirical rank

$r = \hat{R}_i \in (0,1)^{d_x}$ observed in the data. These estimates are then averaged over the data to provide $\hat{\beta}(\mathcal{R})$.

$$\hat{\beta}(\mathcal{R}) = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}(\hat{R}_i)$$

Stored results

Default settings	
e(b)	estimated coefficients
e(N)	number of observations
e(sample)	marks estimation sample
With bootstrapped standard errors	
Matrices	
e(b)	estimated coefficients
e(b bs)	bootstrap estimates
e(reps)	number of nonmissing results
e(bias)	estimated biases
e(se)	estimated standard errors
e(z0)	median biases
e(accel)	estimated accelerations
e(ci normal)	normal-approximation CIs
e(ci percentile)	percentile CIs
e(ci bc)	bias-corrected CIs
e(ci bca)	bias-corrected and accelerated CIs
e(V)	bootstrap variance-covariance matrix
e(V modelbased)	model-based variance
Scalars	
e(N)	sample size
e(N reps)	number of complete replications
e(N misreps)	number of incomplete replications
e(N strata)	number of strata
e(N clust)	number of clusters
e(level)	confidence level for bootstrap CIs
e(bs version)	version for bootstrap results
e(rank)	rank of e(V)

NLSY Illustration

This example uses the dataset of Card (1995) and Kling (2001), which can be loaded into Stata easily by the following commands:

```
net from http://www.stata-press.com/data/musr
net install musr
net get musr
use mus06klingdata
```

The data include 3,010 men aged 24-34 from the National Longitudinal Survey (NLS) Young Men sample.

The outcome of interest is wage76, the worker's log wage earnings in 1976. Following prior research, one includes exogenous controls for age, race, worker's parents' characteristics, region and urban-rural status. The basic endogenous regressor is grade76, the worker's completed years of schooling in 1976. The derived endogenous variables are experience and its square, experience being measured as the individual's age minus years of education. The regression model is

$$\text{wage76}_i = Z_{1i}B_i^z + B_{i1}^x\text{grade76}_i + B_{i2}^x\text{exp76}_i + B_{i3}^x\text{expsq76}_i + e_i$$

where Z_{1i} is a vector of included exogenous covariates and B_i^z the corresponding vector of random coefficients, including a constant. The endogenous random coefficients are notated by B_{i1}^x, measuring the return to schooling, and (B_{i2}^x, B_{i3}^x) measuring returns to experience. The excluded instrumental variables are proximity (while in high school) to a 4-year college: col4pub, col4, and col4m. The first stage model is

$$\text{grade76}_i = Z_{1i}\theta + \gamma_1\text{col4pub}_i + \gamma_2\text{col4}_i + \gamma_3\text{col4m}_i + \tilde{e}_i$$

where γ are coefficients on the excluded exogenous variables. For 2SLS, these coefficients are homogeneous across individuals. For IVCRC, the coefficients and the residual are quantile-specific.

Summary statistic of interest are

```
. ** Endogenous variables and instruments **
. sum wage76 exp76 expsq76 col4pub mcol4 col4

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      wage76 |      3010    1.656664     .443798          0     3.1797
       exp76 |      3010    8.856146    4.141672          0         23
     expsq76 |      3010    .9557907    .8461831          0       5.29
     col4pub |      3010     .492691    .5000296          0          1
       mcol4 |      3010    .6777409    .4674193          0          1
-------------+--------------------------------------------------------
        col4 |      3010    .6820598    .4657535          0          1
```

Running OLS and two stage least squares (2SLS) replicates well known results, suppressing some of the table output:

```
. ** OLS **
. reg wage76 grade76 exp76 expsq76 black south smsa66 smsa76 momdad14 daded momed
famed1-famed8 reg1-reg8

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 26,  2983) =   50.09
       Model |  180.114315    26  6.92747364           Prob > F      =  0.0000
```

```
    Residual |  412.528302  2983  .138293095           R-squared     =  0.3039
-------------+------------------------------           Adj R-squared =  0.2979
       Total |  592.642616  3009  .196956669           Root MSE      =  .37188

------------------------------------------------------------------------------
      wage76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     grade76 |   .0725524   .0036962    19.63   0.000      .065305    .0797998
       exp76 |   .0844427   .0066747    12.65   0.000     .0713551    .0975302
     expsq76 |  -.2289176   .0318981    -7.18   0.000    -.2914621   -.1663731

...
       _cons |   .0973938   .0856921     1.14   0.256    -.0706278    .2654155
------------------------------------------------------------------------------
```

And for IV

```
. ** Manual two-step IV **
. reg grade76 col4pub mcol4 col4 black south smsa66 smsa76 momdad14 daded momed
famed1-famed8 reg1-reg8

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 26,  2983) =   47.88
       Model |  6349.26586    26  244.202533           Prob > F      =  0.0000
    Residual |  15212.8142  2983  5.09983715           R-squared     =  0.2945
-------------+------------------------------           Adj R-squared =  0.2883
       Total |  21562.0801  3009  7.16586243           Root MSE      =  2.2583

------------------------------------------------------------------------------
     grade76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     col4pub |   .2455924   .1158248     2.12   0.034     .0184877     .472697
       mcol4 |  -.4014031   .6373139    -0.63   0.529    -1.651022    .8482161
        col4 |    .600018   .6364568     0.94   0.346    -.6479208    1.847957
...
       _cons |   8.722282   .3641404    23.95   0.000     8.008291    9.436274
------------------------------------------------------------------------------

.
. predict eduhat, xb

. gen experhat = age76 - eduhat
. gen experhatsq = (age76 - eduhat)^2

. reg wage76 eduhat experhat experhatsq black south smsa66 smsa76 momdad14 daded momed
famed1-famed8 reg1-reg8

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 26,  2983) =   43.72
       Model |  163.518419    26  6.28916996           Prob > F      =  0.0000
    Residual |  429.124197  2983  .143856586           R-squared     =  0.2759
-------------+------------------------------           Adj R-squared =  0.2696
       Total |  592.642616  3009  .196956669           Root MSE      =  .37928

------------------------------------------------------------------------------
      wage76 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      eduhat |   .1507692   .0400995     3.76   0.000     .0721438    .2293946
    experhat |   .1524024   .0169069     9.01   0.000      .119252    .1855528
  experhatsq |  -.0036262   .0005387    -6.73   0.000    -.0046824   -.0025701
...
       _cons |   -1.49143   .3846996    -3.88   0.000    -2.245733   -.7371265
------------------------------------------------------------------------------
```

The OLS estimates of returns to schooling are much smaller than IV estimates, 0.07 compared to 0.15, though the IV estimates are less precise. An additional year of experience is worth about as much as an additional year of schooling in both the OLS and 2SLS results. Substituting age for experience disciplines the estimated return to schooling in both cases. The IVCRC estimate of the return to schooling lies in between the two extremes at 0.099:

```
.
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
reg1-reg8 (grade76 = col4pub mcol4 col4), dendog(exp76 expsq76)

Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |   .0994082
       exp76 |   .0956955
     expsq76 |  -.3169567
```

```
...
       _cons |  -.1997724
------------------------------------------------------------------------------
```

where the given command highlights that default settings do not provide bootstrapped standard errors. These can be obtained by specifying option “bootstrap()” as follows:

```
. *Standard errors with bootstrap
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
 reg1-reg8 (grade76 = col4pub mcol4 col4),  boot(reps(25) dendog(exp76 expsq76)

xxxxxxxxxxx
Bootstrap results                               Number of obs      =      3010
                                                Replications       =        14

------------------------------------------------------------------------------
             |    Observed               Bootstrap
      wage76 |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
     grade76 |   .09940817  -.0172084   .01949567     .051482   .1143454   (P)
       exp76 |   .09569555  -.0079764   .00872941    .0709284   .1026235   (P)
     expsq76 |  -.31695668   .0311131   .03110221   -.3321232  -.2135848   (P)
...
       _cons |  -.19977245   .2313945   .22656752   -.3717846   .4840544   (P)
------------------------------------------------------------------------------
(P)    percentile confidence interval
Note: one or more parameters could not be estimated in 11 bootstrap replicates;
      standard-error estimates include only complete replications.
```

Users can test sensitivity of results with respect to the bandwidth choice, choosing one alternative or a list of values for comparison:

```
. *Specifying bandwidth or multiple bandwidths
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
reg1-reg8 (grade76 = col4pub mcol4 col4), bandwidth(.15) dendog(exp76 expsq76)

Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
```

```
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |   .0752979
       exp76 |   .0918175
     expsq76 |  -.2893771
...
       _cons |   .0662526
------------------------------------------------------------------------------
```

Similarly, sensitivity to the choice of kernel function can be examined. E.g. choosing the epanechnikov rather than box kernel:

```
. *Choosing different kernel functions
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
reg1-reg8 (grade76 = col4pub mcol4 col4), kernel(epanechnikov) dendog(exp76 expsq76)

Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |   .1096617
       exp76 |    .096535
     expsq76 |   -.320385
...
       _cons |  -.2995878
------------------------------------------------------------------------------
```

Practitioners can inspect the role numerical grid in computing $\hat{\beta}$. The default set for the numerical integral is $(0.01, 0.02, ..., 0.99)$, and an alternative set on $(0, 1)$ may specified by the user via "integral()":

```
. *Adjusting the integral set
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
reg1-reg8 (grade76 = col4pub mcol4 col4), integral(.2(.01).8) dendog(exp76 expsq76)
```

```
Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |   .1119626
       exp76 |   .1205929
     expsq76 |   -.408651
...
       _cons |  -.5323113
------------------------------------------------------------------------------
```

Computing the rank statistic from quantile regressions is the most intensive step of the module. One can adjust the precision of the computation by increasing or decreasing the number of quantile points to compute. For example,

```
. *Adjusting the number of quantile points
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8
reg1-reg8 (grade76 = col4pub mcol4 col4), quantiles(35) dendog(exp76 expsq76)

Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |   .1008693
       exp76 |   .0961102
     expsq76 |  -.3179172
...
       _cons |    -.23691
------------------------------------------------------------------------------
```

If the user specifies disjoint subsets together with “report”, then the expected return to schooling may be computed over each subset grid as well as overall:

```
. *Integral subsets and reporting them
. ivcrc wage76 black south smsa66 smsa76 momdad14 daded momed famed1-famed8 reg1-reg8
 (grade76 = col4pub mcol4 col4), integral(.01(.01).25, .26(.01).75, .76(.01).99, report)
 dendog(exp76 expsq76)

Default setting is no standard errors
------------------------------------------------------------------------------
      wage76 |      Coef.
-------------+----------------------------------------------------------------
Estimates    |
     grade76 |    .099476
       exp76 |   .0959875
     expsq76 |  -.3179154
...
       _cons |  -.2034218
-------------+----------------------------------------------------------------
subset1      |
     grade76 |   .0832083
       exp76 |   .0249418
     expsq76 |  -.0070813
...
       _cons |   .6375806
-------------+----------------------------------------------------------------
subset2      |
     grade76 |   .1060427
       exp76 |   .1242919
     expsq76 |  -.4138323
...
       _cons |   -.555365
-------------+----------------------------------------------------------------
subset3      |
     grade76 |    .102461
       exp76 |   .1098217
     expsq76 |  -.4379193
...

       _cons |  -.3311975
------------------------------------------------------------------------------
```

Other Notes

1. Mata program and numerical integral

   The stata module uses mata to decipher the contents of the integral() option. The module treats the user's input into integral() as a string, and passes this to the mata program gridparse. Gridparse then does three things. First, if the user specifies more than one set over which to compute the integral, then gridparse organizes these subsets as number lists so that stata can work over each grid. Second, gridparse constructs the proper (lebesgue) weight vectors needed to average coefficients over the subsets. Last, gridparse checks that the grid is compatible (non-overlapping ascending subsets), and passes formating information to stata if the user wants to display results for each subset by specifying the "report" option within integral().

2. Fewer than 1600 grid points per integral subset (Stata 12)

   The integral (and bandwidth) options are built on Stata's number list, or "numlist" syntax. In stata 12, the maximum number of elements in a numlist was 1,600, but in Stata 14 it is higher (2,500). The module is constrained by this limitation, in that each subset specified in the integral() option must contain no more than 1,600 (or more, depending on the user's stata version) or stata will produce an error message. This can be side stepped by breaking large subsets into a few smaller ones.

3. Saving conditional rank statistic

   User's may save the estimated conditional rank statistic to the using dataset by specifying the generate() option and providing a variable name. However, the generate() option is ignored when bootstrapping to obtain standard errors, since each replication has its own estimated rank data.

4. Coefficient e(b) matrix and postestimation

   The module is built to use stata's eclass features, which relies on a storage matrix e(b). When estimates are displayed, standard errors computed, or postestimation predictions made these are based off the contents of e(b). When using postestimation commands like "predict", while also specifying that either multiple bandwidth or multiple subset coefficient vectors be reported, the module places the main estimates into e(b) but places the subset/bandwidth estimates there as well. Thus, "predict" will not know what to do with the extra coefficients. Users must remove the "report" option and specify only one bandwidth if using postestimation commands.