cfregress — Control-function linear regression+

+This command is part of StataNow.

Description

`cfregress` fits linear models with endogenous regressors using control functions. Endogenous variables are first modeled as a function of instruments using linear, probit, fractional probit, or Poisson regression. The residuals, or generalized residuals, from these first-stage regressions are then included in the main equation as control functions to make regression estimates robust to endogeneity.

Quick start

Control function estimation of a linear regression of `y1` on `x` and endogenous regressor `y2` that is instrumented by `z`

```
cfregress y1 x (y2 = z)
```

Same as above, but with two endogenous regressors and two instruments

```
cfregress y1 x (y2 y3 = z1 z2)
```

Same as above, but use `z3` as an additional instrument for `y3`

```
cfregress y1 x (y2 = z1 z2) (y3 = z1 z2 z3)
```

Model the first stage for binary endogenous regressor `y4` using probit regression

```
cfregress y1 x (y2 = z1 z2) (y4 = z1 z2 z3, probit)
```

Include an interaction term between `w` and the control function of `y2` in the main equation

```
cfregress y1 x (y2 = z, interact(w))
```

Include an interaction term between the control functions of `y2` and `y3`

```
cfregress y1 x (y2 = z1 z2) (y3 = z1 z2), cfinteract
```

Include `w` in the main equation for `y1` but not in the first stage

```
cfregress y1 x (y2 = z), mainonly(w)
```

Include an endogenous interaction term between `w` and `y2`, and control for its endogeneity by including an interaction term between `w` and the control function of `y2`

```
cfregress y1 x w (y2 = z, interact(w)), mainonly(c.y2#c.w)
```

Menu

Statistics > Endogenous covariates > Control-function linear regression

Syntax

cfregress *depvar* [*indepvars*] (*varlist*$_{\text{en1}}$ = *varlist*$_{\text{iv1}}$ [, *cfopts*])
[(*varlist*$_{\text{en2}}$ = *varlist*$_{\text{iv2}}$ [, *cfopts*]) . . .] [*if*] [*in*] [*weight*] [, *options*]

cfopts	Description
Model	
linear	model the endogenous variables using linear regression; the default
probit	model the endogenous variables using probit regression
fprobit	model the endogenous variables using fractional probit regression
poisson	model the endogenous variables using Poisson regression
interact(*varlist*$_{\text{int}}$)	interact the variables in *varlist*$_{\text{int}}$ with the control functions

Only one of linear, probit, fprobit, or poisson is allowed in each set of parentheses.

options	Description
Model	
mainonly(*varlist*$_{\text{m}}$)	include the variables in *varlist*$_{\text{m}}$ as exogenous variables in the main equation but not in the first-stage equations
cfinteract	include interactions between control functions when there are multiple endogenous variables
noconstant	suppress constant term
hascons	has user-supplied constant
SE/Robust	
vce(*vcetype*)	*vcetype* may be conventional, robust, cluster *clustvar*, bootstrap, jackknife, or hac *hacspec*
Reporting	
level	set confidence level; default is level(95)
first	report first-stage regressions
noheader	display only the coefficient table
eform[(*string*)]	report exponentiated coefficients and, optionally, label as *string*
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
coeflegend	display legend instead of statistics

indepvars, *varlist*$_{\text{en}\cdot}$, *varlist*$_{\text{iv}\cdot}$, *varlist*$_{\text{int}}$, and *varlist*$_{\text{m}}$ may contain factor variables; see **[U] 11.4.3 Factor variables**.

depvars, *indepvars*, *varlist*$_{\text{en}\cdot}$, *varlist*$_{\text{iv}\cdot}$, *varlist*$_{\text{int}}$, and *varlist*$_{\text{m}}$ may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

bootstrap, by, collect, jackknife, rolling, and statsby are allowed; see **[U] 11.1.10 Prefix commands**.

Weights are not allowed with the bootstrap prefix; see **[R] bootstrap**.

aweights are not allowed with the jackknife prefix; see **[R] jackknife**.

aweights, fweights, iweights, and pweights are allowed; see **[U] 11.1.6 weight**.

coeflegend does not appear in the dialog box.

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

Options

Model

`linear`, `probit`, `fprobit`, and `poisson` specify which regression model is used for the first-stage model. A different model can be specified for each set of parentheses.

`linear`, the default, specifies a linear regression model.

`probit` specifies a probit regression model. Endogenous variables must be coded as 0/1.

`fprobit` specifies a fractional probit regression model. Endogenous variables must take values in $[0, 1]$.

`poisson` specifies a Poisson regression model. Endogenous variables must take nonnegative values.

`interact(`*varlist*$_{\text{int}}$`)` includes in the main regression an interaction term between each variable in *varlist*$_{\text{int}}$ and the control functions associated with the current set of parentheses. Variables are treated as continuous by default.

`mainonly(`*varlist*$_{\text{m}}$`)` includes the variables in *varlist*$_{\text{m}}$ as exogenous variables in the main regression but excludes them from the first-stage regressions.

`cfinteract` specifies that all interactions between control functions be included in the main regression. If there is only one endogenous regressor, and thus only one control function, the option has no effect.

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

`hascons` indicates that a user-defined constant or its equivalent is specified among the independent variables.

SE/Robust

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

`vce(conventional)`, the default, requests conventional standard errors appropriate under homoskedasticity and no autocorrelation.

`vce(hac` *hacspec*`)` requests a heteroskedasticity- and autocorrelation-consistent (HAC) variance–covariance matrix. The full syntax of *hacspec* is one of the following:

`vce(hac` *kernel* [#]`)` requests a HAC variance–covariance matrix using the specified kernel (see below) with optional # lags. The bandwidth of a kernel is equal to # + 1. If # is not specified, a kernel with $N - 2$ lags is used, where N is the sample size.

`vce(hac` *kernel* `opt` [#]`)` requests a HAC variance–covariance matrix using the specified kernel (see below), and the lag order is selected using Newey and West's (1994) optimal lag-selection algorithm. # is an optional tuning parameter that affects the lag order selected; see the discussion in *Methods and formulas* of [R] **ivregress**.

kernel may be one of the following:

`bartlett` or `nwest` requests the Bartlett (Newey–West) kernel.

`parzen` or `gallant` requests the Parzen (Gallant 1987) kernel.

`quadraticspectral` or `andrews` requests the quadratic spectral (Andrews 1991) kernel.

Reporting

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

`first` requests that the results of first-stage regressions be displayed.

`noheader` suppresses the display of the summary statistics at the top of the output, displaying only the coefficient table.

`eform` and `eform(`*string*`)` specify that the coefficient table be displayed in exponentiated form and that `exp(b)` and *string*, respectively, be used to label the exponentiated coefficients in the table. Standard errors and confidence intervals are also transformed.

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**.

The following option is available with `cfregress` but is not shown in the dialog box:

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

Remarks and examples

`cfregress` fits linear models with endogenous regressors by estimating one or more control functions and including them in the main regression equation. These control functions are estimated as the residuals, or generalized residuals, of first-stage regressions.

Control-function methods are closely related to standard instrumental-variables (IV) methods and in the simplest cases produce the same regression estimates. However, control-function methods allow for more flexibility than comparable IV methods. Wooldridge (2015) gives an overview of control-function regression methods.

The main equation in the model fit by `cfregress` is

$$y_{i0} = \mathbf{y}_i\boldsymbol{\beta}_1 + \mathbf{x}_i\boldsymbol{\beta}_2 + \mathbf{w}_i\boldsymbol{\beta}_3 + u_i \tag{1}$$

where y_{i0} is the dependent variable for the ith observation; $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$ is a row vector of p endogenous regressors; $\mathbf{x}_i$ is a row vector of exogenous regressors to be included in the main equation and in first-stage regressions; $\mathbf{w}_i$ is a row vector of exogenous regressors to be included only in the main equation; $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ are vectors of coefficients; and u_i is an error term whose conditional mean is thought to depend on the endogenous variables $\mathbf{y}_i$.

We assume the existence of a set of exogenous instruments for each endogenous regressor. These sets of instruments can be the same across endogenous regressors, or they can be different. Let $\mathbf{z}_i^k$ be the vector containing the instruments for endogenous regressor y_{ik}, and let $\mathbf{z}_i = (\mathbf{z}_i^1, \mathbf{z}_i^2, \dots, \mathbf{z}_i^p)'$ be the vector containing the instruments for all endogenous regressors in the model.

The main equation is similar to those fit by linear IV methods. However, the control-function approach imposes additional structure on the model in that the endogeneity in the error term u_i is explicitly modeled. Specifically, we assume

$$\begin{aligned} E(u_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i) &= E(u_i|\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i) \\ &= \nu_i\boldsymbol{\rho} + h(\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)'\boldsymbol{\rho}_h \end{aligned} \tag{2}$$

Here $\nu_i = (\nu_{i1}, \nu_{i2}, \dots, \nu_{ip})'$ is a row vector of control functions, one for each endogenous variable, and $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_p)$ is a vector of coefficients. $h(\cdot)$ is a known vector-valued function and can include, for our purposes, interactions among the control functions in ν_i, as well as between the control functions and the exogenous or endogenous variables. $\boldsymbol{\rho}_h$ is a set of associated parameters.

For example, suppose that we have one endogenous variable y_{i1} and two instruments z_{i1}^1 and z_{i2}^1 and that $\mathbf{x}_i$ and $\mathbf{w}_i$ are empty. $h(\boldsymbol{\nu}_i, y_{i1}, \mathbf{z}_i^1)$ might take the form $(\nu_{i1} z_{i1}^1, \nu_{i1} z_{i2}^1)$. Combining (1) and (2) and specifying that $\epsilon_i = u_i - E(u_i|\mathbf{y}_i, \mathbf{z}_i)$, we can write

$$y_{i0} = y_{i1}\beta_1 + \nu_{i1}\rho_1 + \nu_{i1} z_{i1}^1 \rho_{h1} + \nu_{i1} z_{i2}^1 \rho_{h2} + \epsilon_i$$

When $h(\cdot) \equiv 0$ and all first-stage models are linear, control-function estimates of the coefficients of the main equation are numerically equivalent to two-stage least-squares IV estimates of the same main equation with the same instruments.

▷ Example 1: Single endogenous regressor, linear first stage

In practice, control functions are not observed but rather estimated. Specifically, the residuals or generalized residuals produced in first-stage regressions serve as control functions. We can model the endogenous variable y_{i1} by the linear regression

$$y_{i1} = \mathbf{x}_i \boldsymbol{\pi}_{11} + \mathbf{z}_i^1 \boldsymbol{\pi}_{12} + \nu_{i1}$$

and use the estimate $\widehat{\nu}_{i1}$ as our control function for y_{i1}.

To illustrate, we revisit example 1 in [R] **ivregress** using census data on housing. We have state data from the 1980 census on the median home value (`hsngval`) and the median monthly gross rent (`rent`). We can model (`rent`) as a function of `hsngval` and the percentage of the population living in urban areas (`pcturban`),

$$\texttt{rent}_i = \beta_0 + \beta_1 \texttt{hsngval}_i + \beta_2 \texttt{pcturban}_i + u_i$$

where i indexes states. We believe that `hsngval` is endogenous; thus, we instrument it using the state's median family income (`faminc`) and census region (`region`).

We can re-create the ivregress 2sls estimates for this model using cfregress. Here, however, we rescale hsngval and faminc to be in thousands of dollars so that they are on a scale similar to rent:

```
. use https://www.stata-press.com/data/r18/hsng
(1980 Census housing data)

. replace hsngval = hsngval/1000
variable hsngval was long now double
(50 real changes made)

. replace faminc = faminc/1000
variable faminc was long now double
(50 real changes made)

. cfregress rent pcturban (hsngval = faminc i.region)

Control-function linear regression                     Number of obs =      50
                                                       Wald chi2(2)  =   90.76
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.5989
                                                       Root MSE      = 22.1656

Endogenous variable model:
    Linear: hsngval

------------------------------------------------------------------------------
        rent | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
rent         |
     hsngval |   2.239833   .3284392     6.82   0.000     1.596104    2.883562
    pcturban |    .081516   .2987652     0.27   0.785    -.504053     .667085
       _cons |   120.7065   15.22839     7.93   0.000     90.85942    150.5536
-------------+----------------------------------------------------------------
e.rent       |
  cf(hsngval)|  -1.588908   .4333422    -3.67   0.000    -2.438243   -.7395726
------------------------------------------------------------------------------
Instruments for hsngval: faminc 2.region 3.region 4.region
```

Accounting for scaling, these estimates are identical to comparable estimates produced by ivregress 2sls, but cfregress also includes an estimate of the coefficient on the control function, reported as cf(hsngval). Here e.rent denotes the model used for u_i, the error term in the main equation for the dependent variable, rent. This error term is modeled as a function of the control functions and, in some cases, other interaction terms involving them. In our example, a test of the hypothesis that the coefficient on cf(hsngval) is different from zero can be interpreted as a test of the endogeneity of hsngval.

We may suspect, however, that our model for u_i in the previous example is misspecified. We can add an interaction term between the control function and `faminc` to this model by using the `interact(faminc)` option:

```
. cfregress rent pcturban (hsngval = faminc i.region, interact(faminc))
Control-function linear regression                     Number of obs =      50
                                                       Wald chi2(2)  =   95.16
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.5945
                                                       Root MSE      = 22.2851
Endogenous variable model:
    Linear: hsngval

------------------------------------------------------------------------------
        rent | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
rent         |
     hsngval |   2.155381   .3437284     6.27   0.000     1.481686    2.829076
    pcturban |   .4794597   .2362242     2.03   0.042     .0164688    .9424506
       _cons |   98.15909   13.86958     7.08   0.000     70.97521     125.343
-------------+----------------------------------------------------------------
e.rent       |
  cf(hsngval)|   10.66765   3.619442     2.95   0.003     3.573673    17.76163
             |
  cf(hsngval)|
      faminc |  -.5610651   .1743049    -3.22   0.001    -.9026965   -.2194338
------------------------------------------------------------------------------
Instruments for hsngval: faminc 2.region 3.region 4.region
```

The coefficient on the endogenous variable, `hsngval`, is slightly different but not substantially changed. However, the coefficient on `pcturban` is now noticeably larger, and there is evidence it is different from zero. The coefficient on the control function `cf(hsngval)` has changed sign, and there is evidence that the coefficient on the interaction term is also relevant in the model, suggesting that including it in the model for the error term is appropriate. In this case, a joint test of `cf(hsngval)` and `cf(hsngval)#faminc` is equivalent to a test of the endogeneity of `hsngval`.

We can perform this test using the postestimation command `estat endogenous`.

```
. estat endogenous
Tests of endogeneity
H0: Variables are exogenous
 ( 1)  [e.rent]cf(hsngval) = 0
 ( 2)  [e.rent]cf(hsngval)#c.faminc = 0
           chi2(  2) =   15.30
         Prob > chi2 =    0.0005
```

◁

We note here that as long as the instruments are valid, misspecification of the endogeneity in the error term, such as by using a two-stage least-squares IV estimator when the data-generating process has $h(u_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i) \neq 0$, will not affect the consistency of the regression estimates. However, it may lead to biased estimates or invalid inference in small samples. Kim and Petrin (2011) discuss issues related to the specification of the endogeneity in the error term.

▷ Example 2: Endogenous variables entering as interactions

Oftentimes, we have a model with a single endogenous regressor, y_{i1}, that appears in the main equation interacted with an exogenous variable x_{i1},

$$y_{i0} = y_{i1}\boldsymbol{\beta}_1 + y_{i1}x_{i1}\boldsymbol{\beta}_2 + \mathbf{x}_i\boldsymbol{\beta}_3 + u_i$$

In these cases, it is natural to model u_i as a linear function of the control function for y_{i1}, ν_1, and the interaction term $\nu_1 x_{i1}$,

$$E(u_i|y_{i1}, \mathbf{x}_i, \mathbf{z}_i^1) = \rho_1\nu_1 + \rho_2\nu_1 x_{i1}$$

We can add an interaction term to the main equation and, at the same time, model u_i using the `mainonly()` and `interact()` options.

Returning to the housing value model in the previous example, suppose that we want to include `faminc` as an exogenous variable and to include an interaction term between it and the endogenous regressor `hsngval` in the main equation for `rent`. At the same time, we want to control for endogeneity by including the interaction term between the control function of `hsngval` and `faminc`. We can type

```
. cfregress rent pcturban faminc (hsngval = i.region, interact(faminc)),
> mainonly(c.hsngval#c.faminc)

Control-function linear regression                     Number of obs =      50
                                                       Wald chi2(4)  =  169.37
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.7694
                                                       Root MSE      = 16.8055

Endogenous variable model:
    Linear: hsngval

------------------------------------------------------------------------------
        rent | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
rent         |
     hsngval |  -.1299135    1.66665    -0.08   0.938    -3.396487     3.13666
             |
   c.hsngval#|
    c.faminc |   .0744869    .078061     0.95   0.340    -.0785099    .2274836
             |
    pcturban |    .405831   .2078664     1.95   0.051    -.0015798    .8132417
      faminc |   .7928874   3.834329     0.21   0.836    -6.722259    8.308034
       _cons |   125.9643   73.74242     1.71   0.088    -18.56818    270.4968
-------------+----------------------------------------------------------------
e.rent       |
 cf(hsngval) |   7.225214   2.793552     2.59   0.010     1.749952    12.70048
             |
 cf(hsngval) |
      faminc |   -.374434   .1321422    -2.83   0.005     -.633428   -.1154401
------------------------------------------------------------------------------
Instruments for hsngval: 2.region 3.region 4.region
```

By specifying `mainonly(c.hsngval#c.faminc)`, we request that the interaction term is included in the main equation but not in the first-stage regression for `hsngval`. By specifying `interact(faminc)`, we request that the interaction term `cf(hsngval)#faminc` be included to control for endogeneity.

◁

▷ Example 3: Binary endogenous regressor, probit first stage

Because control-function methods involve explicitly specifying first-stage models for the endogenous variables, the first stage need not be restricted to linear regression. `cfregress` allows for probit, fractional probit, and Poisson regression in the first-stage models for the endogenous variables. This flexibility means that a surprising variety of models can be estimated using `cfregress`.

For instance, example 2 in [CAUSAL] **etregress** uses a control-function estimator to estimate the effect of having health insurance (`ins`) on the log of prescription drug expenditure (`lndrug`). The endogenous binary treatment variable, `ins`, is instrumented using marital status (`married`) and employment status (`work`). We can reproduce the estimates in the example using `cfregress` with the `probit`, `interact()`, and `mainonly()` options:

```
. use https://www.stata-press.com/data/r18/drugexp
(Prescription drug expenditures)

. cfregress lndrug age lninc (ins = i.married i.work, probit interact(i.ins)),
> mainonly(i.chron) vce(robust)

Control-function linear regression                Number of obs =    6,000
                                                  Wald chi2(4)  = 1973.78
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.2432
                                                  Root MSE      =  1.2172

Endogenous variable model:
    Probit: 1.ins

------------------------------------------------------------------------------
             |               Robust
      lndrug | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
lndrug       |
       1.ins |  -.8598836   .3483648    -2.47   0.014    -1.542666   -.1771011
     1.chron |   .4671725   .0319731    14.61   0.000     .4045064    .5298387
         age |   .1021359     .00292    34.98   0.000     .0964128    .1078589
       lninc |   .0550672   .0225036     2.45   0.014     .0109609    .0991735
       _cons |   1.665539   .2527527     6.59   0.000     1.170153    2.160925
-------------+----------------------------------------------------------------
e.lndrug     |
   cf(1.ins) |   .5252243    .226367     2.32   0.020     .0815532    .9688954
             |
 cf(1.ins)#ins |
           0 |          0  (omitted)
           1 |   .2702095   .2585099     1.05   0.296    -.2364605    .7768796
------------------------------------------------------------------------------
Instruments for 1.ins: 1.married 1.work
```

Here the first stage is specified as a probit model and an interaction term is included so that the main error term depends on the value of the treatment variable. We specify the covariate `chron` (whether the individual has a chronic health condition) as present in the main equation but not in the treatment model by using the `mainonly()` option. The main equation, accounting for the specification of endogeneity, thus takes the form

$$\begin{aligned}\texttt{lndrug} = \beta_0 + \beta_1\texttt{ins} + \beta_2\texttt{age} + \beta_3\texttt{lninc} + \beta_4\texttt{chron} \\ + \rho_1\texttt{cf(1.ins)} + \rho_{h1}\texttt{cf(1.ins)\#ins} + \epsilon\end{aligned}$$

The first-stage model is a standard probit model with `ins` as the left-hand side variable and `married`, `work`, `age`, and `lninc` as right-hand side variables. The estimated control function `cf(1.ins)` is the generalized error from this first-stage probit regression, as defined in *Methods and formulas*.

The output allows us to assess the way the endogeneity of ins has implicitly been specified by the treatment-regression model. The model allows the main error term to be correlated with the treatment model error conditional on the value of the treatment variable, which implies that the conditional mean of the main error depends on the interaction of cf(1.ins) and ins. The estimated coefficient on cf(1.ins)#ins does not, as it turns out, give evidence that this is the case.

Given this absence of evidence for a treatment-specific error term, we may wish to estimate a regression that does not include the interaction. This is not possible with etregress, but in cfregress we can simply drop the interact() option.

```
. cfregress lndrug age lninc (ins = i.married i.work, probit),
> mainonly(i.chron) vce(robust)

Control-function linear regression                     Number of obs =   6,000
                                                       Wald chi2(4)  = 2833.77
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.2393
                                                       Root MSE      =  1.2203

Endogenous variable model:
    Probit: 1.ins

------------------------------------------------------------------------------
             |               Robust
      lndrug | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
lndrug       |
       1.ins |  -.8992025   .3399829    -2.64   0.008    -1.565557   -.2328483
     1.chron |   .4675479   .0319717    14.62   0.000     .4048845    .5302113
         age |   .1011597   .0027163    37.24   0.000     .0958359    .1064836
       lninc |   .0505756   .0217621     2.32   0.020     .0079228    .0932285
       _cons |   1.827957   .1784883    10.24   0.000     1.478126    2.177787
-------------+----------------------------------------------------------------
e.lndrug     |
    cf(1.ins)|   .6157838   .1991464     3.09   0.002      .225464    1.006104
------------------------------------------------------------------------------
Instruments for 1.ins: 1.married 1.work
```

Here we find a slightly more extreme main effect of ins and slightly more precise estimates of each of the coefficients after dropping the interaction term.

◁

Because we have explicit control of the model for the conditional expectation of the error term, we can not only treat other kinds of models as special cases but also customize these special cases for our setting, as in the preceding binary treatment example. Similarly, control function regression can be used to flexibly model correlated random coefficients in linear models, as outlined in Wooldridge (2015).

cfregress also allows for multiple endogenous regressors, each of which can have a different first-stage model and set of instruments.

Stored results

`cfregress` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(k_endog)`	number of endogenous variables
`e(df_m)`	model degrees of freedom
`e(rmse)`	root mean squared error
`e(r2)`	R^2
`e(chi2)`	χ^2
`e(p)`	p-value for model test
`e(N_clust)`	number of clusters
`e(hac_lag)`	HAC lag
`e(rank)`	rank of `e(V)`

Macros

`e(cmd)`	`cfregress`
`e(cmdline)`	command as typed
`e(depvar)`	name of dependent variable
`e(endog)`	names of endogenous variables
`e(exog)`	names of exogenous variables
`e(exog_main)`	names of exogenous variables in main equation only
`e(constant)`	`noconstant` or `hasconstant`, if specified
`e(wtype)`	weight type
`e(wexp)`	weight expression
`e(modeltypes)`	model specification (`linear`, `probit`, etc.) for each endogenous regressor
`e(cfinteract)`	`cfinteract`, if specified
`e(title)`	title in estimation output
`e(clustvar)`	name of cluster variable
`e(hac_kernel)`	HAC kernel
`e(vce)`	*vcetype* specified in `vce()`
`e(vcetype)`	title used to label Std. err.
`e(exogr)`	exogenous regressors
`e(properties)`	`b V`
`e(estat_cmd)`	program used to implement `estat`
`e(predict)`	program used to implement `predict`
`e(footnote)`	program used to implement footnote display
`e(marginsok)`	predictions allowed by `margins`
`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

As discussed in *Remarks and examples*, the main equation estimated by `cfregress` has the form

$$y_{i0} = \mathbf{y}_i\boldsymbol{\beta}_1 + \mathbf{x}_i\boldsymbol{\beta}_2 + \mathbf{w}_i\boldsymbol{\beta}_3 + u_i$$

where y_{i0} is the dependent variable for the ith observation; $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$ is a row vector of p endogenous regressors; $\mathbf{x}_i$ is a row vector of exogenous regressors to be included in the main equation and in first-stage regressions; $\mathbf{w}_i$ is a row vector of exogenous regressors to be included only in the main equation; $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ are vectors of coefficients; and u_i is an error term whose conditional mean is thought to depend on the endogenous variables $\mathbf{y}_i$.

We also specify first-stage models for each endogenous regressor y_{ik} as a function of the exogenous regressors $\mathbf{x}_i$ and instruments $\mathbf{z}_i^k$. If a linear model is specified for y_{ik}, either using the `linear` option or by default, its first-stage equation has the form

$$y_{ik} = \mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2} + \nu_{ik}$$

where $\boldsymbol{\pi}_{k,1}$ and $\boldsymbol{\pi}_{k,2}$ are coefficients and ν_{ik} is an error term.

If a probit model is specified using the `probit` option, the first-stage model for y_{ik} has the form

$$P(y_{ik} = 1|\mathbf{x}_i, \mathbf{z}_i^k) = \Phi(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})$$

If a fractional probit model is specified using the `fprobit` option, the first-stage model for the conditional mean of y_{ik} can be written as

$$E(y_{ik}|\mathbf{x}_i, \mathbf{z}_i^k) = \Phi(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})$$

Finally, if a Poisson model is specified using the `poisson` option, the first-stage model for the conditional mean of y_{ik} can be written as

$$E(y_{ik}|\mathbf{x}_i, \mathbf{z}_i^k) = \exp(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})$$

For each endogenous variable y_{ik}, a control function $\nu_{ik}(y_{ik}, \mathbf{x}_i, \mathbf{z}_i^k)$ is estimated. In the linear case, this is simply the estimate of the linear error term.

For probit and fractional probit models, it is an estimate at the optimum of the "generalized error", which is equal to the first derivative of the probit log likelihood with respect to $\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2}$,

$$\nu_{ik}(y_{ik}, \mathbf{x}_i, \mathbf{z}_i^k) = y_{ik}\frac{\phi(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})}{\Phi(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})} - (1 - y_{ik})\frac{\phi\{-(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})\}}{\Phi\{-(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})\}}$$

For a Poisson first-stage model, $\nu_{ik}(y_{ik}, \mathbf{x}_i, \mathbf{z}_i^k)$ is equal to the first derivative of the Poisson log likelihood with respect to $\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2}$,

$$\nu_{ik}(y_{ik}, \mathbf{x}_i, \mathbf{z}_i^k) = y_{ik} - \exp(\mathbf{x}_i\boldsymbol{\pi}_{k,1} + \mathbf{z}_i^k\boldsymbol{\pi}_{k,2})$$

We also assume a known form for the endogeneity in u_i. Specifically,

$$\begin{aligned} E(u_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i) &= E(u_i|\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i) \\ &= \nu_i\boldsymbol{\rho} + h(\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)'\boldsymbol{\rho}_h \end{aligned}$$

where $\boldsymbol{\nu}_i = \{\nu_{i1}(y_{i1}, \mathbf{x}_i, \mathbf{z}_i^1), \dots, \nu_{ip}(y_{ip}, \mathbf{x}_i, \mathbf{z}_i^p)\}'$, $h(\cdot)$ is known, $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_p)$, $\boldsymbol{\rho}_h$ is a vector of coefficients associated with the elements of $h(\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)$, and $\mathbf{z}_i = (\mathbf{z}_i^1, \mathbf{z}_i^2, \dots, \mathbf{z}_i^p)'$. Accordingly, we produce estimates of our regression coefficients using a modified main equation of the form

$$y_{i0} = \mathbf{y}_i\boldsymbol{\beta}_1 + \mathbf{x}_i\boldsymbol{\beta}_2 + \mathbf{w}_i\boldsymbol{\beta}_3 + \widehat{\boldsymbol{\nu}}_i\boldsymbol{\rho} + h(\widehat{\boldsymbol{\nu}}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)'\boldsymbol{\rho}_h + \epsilon_i$$

where $\widehat{\boldsymbol{\nu}}_i$ is an estimate of $\boldsymbol{\nu}_i$ computed in first-stage regressions and ϵ_i is an error term. This equation is estimated by ordinary least squares to produce estimates of the coefficients that are appropriately corrected for endogeneity.

Similarly to two-stage least-squares instrumental-variables estimation, the standard errors returned by this two-stage procedure will be incorrect, because ϵ_i will be incorrectly taken as the overall error term, rather than as a component of the true overall error term $u_i = \boldsymbol{\nu}_i\boldsymbol{\rho} + h(\boldsymbol{\nu}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)'\boldsymbol{\rho}_h + \epsilon_i$.

To correct this, the standard errors are computed as if the model was estimated using generalized method of moments (GMM). The GMM specification used to produce standard errors includes a set of moment conditions for the main equation, as well as a set of moment conditions for each of the first-stage models.

The error function for the dependent variable is

$$\begin{aligned}\epsilon_i(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\rho}, \boldsymbol{\rho}_h) = {} & y_{i0} - \mathbf{y}_i\boldsymbol{\beta}_1 - \mathbf{x}_i\boldsymbol{\beta}_2 - \mathbf{w}_i\boldsymbol{\beta}_3 - \boldsymbol{\nu}_i(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\rho} \\ & - h(\boldsymbol{\nu}_i(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i), \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)'\boldsymbol{\rho}_h\end{aligned}$$

It forms a set of moment conditions with associated instruments $\mathbf{y}_i$, $\mathbf{x}_i$, $\mathbf{w}_i$, $\widehat{\boldsymbol{\nu}}_i$, and $h(\widehat{\boldsymbol{\nu}}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)$.

Additionally, each of the control functions $\boldsymbol{\nu}_{ik}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i^k)$ is taken as an error function that forms a set of moment conditions with the exogenous variables $\mathbf{x}_i$ and associated instruments $\mathbf{z}_i^k$.

Together, these moment conditions define an exactly identified model for the purpose of GMM estimation, even if there are more instruments in $\mathbf{z}_i$ than there are endogenous variables (in this sense, it is a method of moments specification). This is because each instrument in the moment conditions is associated with a unique parameter. Because the GMM model is exactly identified, the results are invariant to the choice of the GMM weight matrix.

Acknowledgment

We thank Jeffrey M. Wooldridge of the Department of Economics at Michigan State University for his extensive contributions to the literature on control-function methods.

References

Andrews, D. W. K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59: 817–858. https://doi.org/10.2307/2938229.

Gallant, A. R. 1987. *Nonlinear Statistical Models*. New York: Wiley. https://doi.org/10.1002/9780470316719.

Kim, K., and A. K. Petrin. 2011. A new control function approach for non-parametric regressions with endogenous variables. Working Paper 16679, National Bureau of Economic Research. https://doi.org/10.3386/w16679.

Newey, W. K., and K. D. West. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* 61: 631–653. https://doi.org/10.2307/2297912.

Wooldridge, J. M. 2015. Control function methods in applied econometrics. *Journal of Human Resources* 50: 420–445. https://doi.org/10.3368/jhr.50.2.420.

Also see

[R] **cfregress postestimation** — Postestimation tools for cfregress$^+$

[R] **cfprobit** — Control-function probit regression$^+$

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

[R] **regress** — Linear regression

[CAUSAL] **eteffects** — Endogenous treatment-effects estimation

[CAUSAL] **etregress** — Linear regression with endogenous treatment effects

[U] 20 Estimation and postestimation commands

For suggested citations, see the FAQ on citing Stata documentation.