# Control-function models in Stata

Tom Stringham
Senior Econometrician and Software Developer
StataCorp LLC

2025 Stata Conference
July 31, 2025

# Control Functions

## Outline

```
cfregress y w (x = z1 z2), vce(robust)
```

To discuss:

- The idea behind control functions
- The simple linear case: how it works
- Variations: theory, syntax, example output
- VCE/Standard errors
- Postestimation

# Control Functions

## IV, but with more structure

Regression models often suffer from endogeneity.

- e.g. $y = \beta_0 + \beta_1 x + u$ : we want $\beta_1$, but when $x$ moves in our data, so does $u$

Our workhorses for these models are instrumental-variables (IV) methods.

- Idea: find an *instrument* $z$ that moves $x$ but not $u$. Then "move" $z$ enough to move $x$ one unit, and see what happens to $y$.

Control function (CF) methods are a variation on plain IV.

- Idea: model the part of $x$ that $z$ can*not* explain, call it a *control function*, $v$, then include an estimate $\hat{v}$ in our regression.

# Why use CF if we have IV/2SLS?

Short answer: more flexibility (at the cost of stronger assumptions)

Long answer:

- Built-in tests of endogeneity
- Easy estimation of some correlated random coefficient models
- Simplified handling of endogenous variables entering as interactions
- Exploit discreteness of binary endogenous variables
- And more …
- See Wooldridge (JHR, 2015)

# Control Functions

## They've been there all along

Stata commands that already used CF methods:

- `etregress, cfunction`
- `eteffects`
- `ivprobit, twostep`
- `ivtobit, twostep`
- `ivpoisson cfunction`

The idea of `cfregress` and `cfprobit`: control function regression commands that let users manipulate the CF specification and exploit the distinctive features of control functions, while taking care of standard errors.

Plain linear IV setup with one endogenous regressor (exogenous regressors partialled out):

$$
\begin{aligned}
y &= \beta x + u, \\
x &= Z\pi + v, \\
E(Zu) &= 0, \\
\pi &\neq 0, \\
E(Zv) &= 0.
\end{aligned}
$$

Note that the endogeneity of $x$, $E(xu) \neq 0$, implies that $u$ and $v$ are correlated. Let $\rho = E(uv)/E(v^2)$, and let $\varepsilon = u - \rho v$.

Substituting $u = \rho v + \varepsilon$ into our main equation, we have

$$y = \beta x + \rho v + \varepsilon.$$

We have that $E(x\varepsilon) = E(v\varepsilon) = E(uv) - \rho E(v^2) = 0$, so $x$ is uncorrelated with $\varepsilon$ and we can estimate $\beta$ if we observe $v$.

We do not observe $v$, so in practice we use $\hat{v} = x - Z\hat{\pi}$. Because $\hat{\pi}$ is a consistent estimator of $\pi$, we still get a consistent estimate of $\beta$ (note that $v - \hat{v} = Z(\hat{\pi} - \pi)$).

# Control Functions

## The linear case

$$y = \beta x + \rho \hat{v} + \varepsilon.$$

Note: a test of $\rho = 0$ is a valid test for endogeneity.

Another existing use of control functions: `estat endogenous` after `ivregress 2sls, vce(robust)`.

$$y = \beta x + u,$$
$$x = Z\pi + v.$$

In this linear model, $\hat{\beta}_{CF} = \hat{\beta}_{2SLS}$.

- Intuition: CF uses $x$ along with its first-stage residuals $x - Z\hat{\pi}$ while 2SLS uses fitted values $Z\hat{\pi}$, but both contain the same information about $Z\pi$.
- In other models, there is generally not an IV method using fitted values that is equivalent to CF.

$$y = \beta x + \rho\hat{v} + \varepsilon,$$

Note the above can be rewritten:

$$y = \beta(\hat{x} + \hat{v}) + \rho\hat{v} + \varepsilon,$$

$$y = \beta\hat{x} + (\beta + \rho)\hat{v} + \varepsilon,$$

And since $\hat{x}$ and $\hat{v}$ are orthogonal, we get the same estimate of $\beta$ by running:

$$y = \beta\hat{x} + \textit{error}.$$

Note: $\hat{x}$ is not an estimator of $x$, which is known; it is an estimator of $Z\pi$.

# Control Functions

## The linear case

$$y = \beta x + \rho \hat{v} + \varepsilon.$$

To estimate $\beta$ and $\rho$ in `cfregress`, we use syntax familiar from `ivregress`:

```
cfregress y (xvars = zvars).
```

We can include exogenous variables:

```
cfregress y w1 w2 (xvars = zvars).
```

# Example output

```
. cfregress rent pcturban (hsngval = faminc i.region), vce(robust)
Control-function linear regression                Number of obs =      50
                                                  Wald chi2(2)  =   44.98
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.5989
                                                  Root MSE      = 22.1656

Endogenous variable model:
    Linear: hsngval
```

| rent | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| rent | | | | | | |
| hsngval | .0022398 | .000672 | 3.33 | 0.001 | .0009227 | .0035569 |
| pcturban | .081516 | .4445938 | 0.18 | 0.855 | -.789872 | .9529039 |
| _cons | 120.7065 | 15.25546 | 7.91 | 0.000 | 90.80636 | 150.6067 |
| e.rent | | | | | | |
| cf(hsngval) | -.0015889 | .000806 | -1.97 | 0.049 | -.0031687 | -9.10e-06 |

```
Instruments for hsngval: faminc 2.region 3.region 4.region
```

In a conditional mean sense, CF methods can be thought of as saying

$$E(u|v, Z) = \rho v.$$

But suppose it does not hold.

For example, we could have $E(u|v, Z) = \rho_1 v + \rho_2 vw$, for an exogenous variable $w$. We may think of including an interaction term in our estimating equation:

$$y = \beta_0 + \beta_1 x + \gamma w + \rho_1 v + \rho_2 vw + \eta$$

# Variations

## Nonlinearity of the error in $v$

$$y = \beta_0 + \beta_1 x + \gamma w + \rho_1 v + \rho_2 vw + \eta$$

With this setup, you can show we need $u - \rho_2 vw$ uncorrelated with $Z$. Under our original assumption $E(Zu) = 0$, this means we need $\rho_2 vw$ uncorrelated with $Z$. Because $w$ is part of $Z$, we can get this by assuming $E(v|Z) = 0$.

Or, we can say we don't need $E(Zu) = 0$, but rather $E(Z(u - \rho_2 vw)) = 0$. In other words, we only need $Z$ to be exogenous to whatever is left over after partialling out $vw$. This condition is implied by our CF assumption $E(u|Z, v) = \rho_1 v + \rho_2 vw$.

If we can impose independence then either way is fine, but here we can see that there exist DGPs where the interacted CF approach will give valid results and the regular CF approach won't!

# Variations

## Nonlinearity of the error in $v$

$$y = \beta_0 + \beta_1 x + \gamma w + \rho_1 \hat{v} + \rho_2 \hat{v} w + \eta$$

Command:

```
cfregress y w (x = z1 z2, interact(w))
```

# Example output

```
. cfregress rent pcturban (hsngval = faminc i.region, interact(pcturban)), vce(robust)
Control-function linear regression                  Number of obs =      50
                                                    Wald chi2(2)  =   44.83
                                                    Prob > chi2   =  0.0000
                                                    R-squared     =  0.5574
                                                    Root MSE      = 23.2829

Endogenous variable model:
    Linear: hsngval
```

| rent | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| rent | | | | | | |
| hsngval | .0024082 | .0006391 | 3.77 | 0.000 | .0011556 | .0036608 |
| pcturban | .1459889 | .4308807 | 0.34 | 0.735 | -.6985218 | .9904997 |
| _cons | 108.2288 | 17.36071 | 6.23 | 0.000 | 74.20243 | 142.2552 |
| e.rent | | | | | | |
| cf(hsngval) | .0015522 | .0019371 | 0.80 | 0.423 | -.0022444 | .0053488 |
| cf(hsngval)#pcturban | -.0000419 | .0000236 | -1.78 | 0.075 | -.0000881 | 4.26e-06 |

```
Instruments for hsngval: faminc 2.region 3.region 4.region
```

# Variations

## Models with endogenous regressors entering as interactions

Suppose $x_1$ is endogenous, $z_1$ and $z_2$ are exogenous and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 z_2 + u.$$

The IV way to approach this would be to treat $x_1$ and $x_1 z_2$ as two endogenous regressors that share two instruments $z_1$ and $z_1 z_2$.

```
ivregress 2sls y (x1 c.x1#c.z2 = z1 c.z1#c.z2)
```

# Variations

## Models with endogenous regressors entering as interactions

A control function approach is to model a control function only for $x_1$, with instrument $z_1$, and estimate the regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 z_2 + \hat{v}\rho_1 + error,$$

or even

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 z_2 + \hat{v}\rho_1 + \hat{v} z_2 \rho_2 + error.$$

Commands:

```
cfregress y (x1 = z1), mainonly(c.x1#c.z2)
cfregress y (x1 = z1, interact(z2)), mainonly(c.x1#c.z2)
```

# Variations

## Variables to appear only in the main equation

Note we use the option `mainonly()` to specify a variable that should be treated as exogenous, but should not appear in the first stage.

`ivregress` includes all exogenous variables in the first stage. `cfregress` does too (except for those specified in `mainonly()`), because there is seldom good reason for doing otherwise.

We may be interested in a model with correlated random coefficients:

$$y = \beta_0 + \beta_1(\omega)x + u; \text{ with } \beta_1(\omega) = \beta_1 + \omega.$$

where $\omega$ is a random variable with mean zero. We can write this as

$$y = \beta_0 + \beta_1 x + \omega x + u,$$
$$x = Z\pi + v.$$

We take the error to be $\omega x + u$ and project both $\omega$ and $u$ onto $v$. We then estimate

$$y = \beta_0 + \beta_1 x + \rho_1 x\hat{v} + \rho_2 \hat{v} + error.$$

$$y = \beta_0 + \beta_1 x + \rho_1 x \hat{v} + \rho_2 \hat{v} + error.$$

Command:

```
cfregress y (x = z, interact(x))
```

Conveniently, we can test the heterogeneity of $\beta_1(\omega)$ in $x$, as a test of the null $\rho_1 = 0$.
See Wooldridge (JHR, 2015) for a discussion.

We may have binary $x_1$ and believe it is probit conditional on $Z$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$
$$x_1 = \mathbb{1}(Z\pi_1 + \pi_2 x_2 + v > 0).$$

The CF approach involves estimating

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \hat{r}\rho + \varepsilon,$$

where $\hat{r}$ is the score from the first-stage probit. Under appropriate assumptions, this is valid. The two-step IV approach of plugging in fitted values is generally not.

Command:

```
cfregress y (x1 = z, probit) x2, vce(robust)
```

# Example output

```
. 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
```

| lndrug | Coefficient | Robust 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
```

We can have multiple endogenous regressors and multiple control functions:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
x_1 &= \pi_{10} + \pi_{11} z_1 + \pi_{12} z_2 + v_1 \\
x_2 &= \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + v_2
\end{aligned}
$$

and estimate the equation

$$
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \rho_1 \hat{v}_1 + \rho_2 \hat{v}_2 + \epsilon.
$$

Command:

```
cfregress y (x1 x2 = z1 z2), vce(robust)
```

Note: this is still equivalent to 2SLS despite multiple endogenous variables.

We typically assume:

$$E(u|v_1, v_2, z_1, z_2) = \rho_1 v_1 + \rho_2 v_2.$$

However, this may feel like a strong assumption. We can consider adding an interaction term and estimating:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \rho_1 \hat{v}_1 + \rho_2 \hat{v}_2 + \rho_3 \hat{v}_1 \hat{v}_2 + error.$$

Command:

```
cfregress y (x1 x2 = z1 z2), vce(robust) cfinteract.
```

The `cfinteract` option works using "##" interaction logic, where all combinations of control functions are interacted and included. Note: no effect with only one endogenous variable.

# Example output

```
. cfregress mpg (price foreign = weight length), cfinteract vce(robust)
Control-function linear regression                  Number of obs =     74
                                                    Wald chi2(2)  =  21.70
                                                    Prob > chi2   = 0.0000
                                                    Root MSE      = 9.5680

Endogenous variable models:
    Linear: price foreign
```

| mpg | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| mpg | | | | | | |
| price | .0005727 | .0012578 | 0.46 | 0.649 | -.0018926 | .0030379 |
| foreign | 20.41465 | 7.523141 | 2.71 | 0.007 | 5.669567 | 35.15974 |
| _cons | 12.22066 | 9.600274 | 1.27 | 0.203 | -6.595535 | 31.03685 |
| e.mpg | | | | | | |
| cf(price) | -.0005092 | .0012874 | -0.40 | 0.692 | -.0030324 | .002014 |
| cf(foreign) | -21.18877 | 7.245094 | -2.92 | 0.003 | -35.38889 | -6.988643 |
| cf(foreign)#cf(price) | -.0010821 | .0004913 | -2.20 | 0.028 | -.0020452 | -.0001191 |

```
Instruments for price:   weight length
Instruments for foreign: weight length
```

In `cfregress`, we even allow users to specify endogenous regressors with different instrument sets:

$$
\begin{aligned}
y &= \beta_1 x_1 + \beta_2 x_2 + u, \\
x_1 &= \pi_{11} z_1 + \pi_{12} z_2 + v_1, \\
x_2 &= \pi_{21} z_2 + v_2.
\end{aligned}
$$

Command:

```
cfregress y (x1 = z1 z2) (x2 = z2), vce(robust).
```

# Example output

```
. cfregress mpg (price = weight length) (foreign = length), vce(robust)
Control-function linear regression                    Number of obs =     74
                                                      Wald chi2(2)  =  47.71
                                                      Prob > chi2   = 0.0000
                                                      R-squared     = 0.0035
                                                      Root MSE      = 5.7361

Endogenous variable models:
    Linear: price foreign
```

| mpg | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| mpg | | | | | | |
| price | -.0009205 | .0005543 | -1.66 | 0.097 | -.002007 | .000166 |
| foreign | 13.07771 | 3.601915 | 3.63 | 0.000 | 6.018082 | 20.13733 |
| _cons | 23.0844 | 4.048762 | 5.70 | 0.000 | 15.14897 | 31.01982 |
| e.mpg | | | | | | |
| cf(price) | .0008831 | .0005937 | 1.49 | 0.137 | -.0002805 | .0020467 |
| cf(foreign) | -14.65215 | 3.523867 | -4.16 | 0.000 | -21.5588 | -7.745497 |

```
Instruments for price:  weight length
Instrument for foreign: length
```

We still need the same main control function assumption as in the previous example,

$$E(u|v_1, v_2, z_1, z_2) = \rho_1 v_1 + \rho_2 v_2,$$

but now we are implicitly allowing $z_1$ to be correlated with $v_2$. But note this is an unusual combination of assumptions—usually it is safer to use both instruments for both endogenous variables.

# Standard errors

Standard linear case when we have $y = \beta x + u$:

$$x = Z\pi + v,$$
$$y = \beta x + \rho\hat{v} + error.$$

As will be familiar from 2SLS, the standard errors produced by running the two stages sequentially will be wrong.

One good option is to use option `vce(bootstrap)`.

Ideally, however, we would like an analytic option.

# Standard errors using GMM

Note that it is well known that GMM produces the 2SLS estimator of $\beta$, with appropriate standard errors, when used with the right weights.

$$\sum z_i'(y_i - \beta x_i) = 0,$$
$$\text{with } W = \left(\sum z_i' z_i\right)^{-1}.$$

This works because the GMM objective function then includes a projection matrix $P_Z$ with elements $z_j(\sum z_i' z_i)^{-1} z_k'$, which leads to the familiar estimator $\hat{\beta}_{2SLS} = (X' P_Z X)^{-1} X' P_Z y$.

# Standard errors using GMM

You can write a numerically equivalent exactly-identified GMM system without weights, by taking advantage of the fact that $P_Z P_Z = P_Z$. Specifically, the following sample moment condition gives the same estimates and variance estimator:

$$\sum \hat{x}_i(y_i - \beta x_i) = 0.$$

So, you can run

```
gmm (y - {b}*x), inst(xhat)
```

and get 2SLS estimates and standard errors.

# Standard errors using GMM

Note this condition can be written as $\sum \hat{x}_i(y_i - \beta z_i\pi - \beta(x_i - z_i\pi)) = 0$. We can even write down the following and still get the same estimates and standard errors:

$$\sum \hat{x}_i(y_i - \beta z_i\pi) = 0,$$
$$\sum z_i'(x_i - z_i\pi) = 0.$$

(See Newey, 1984.) Finally, we can introduce the remaining component of the error, $(\rho + \beta)v$, and rearrange:

$$\sum \hat{x}_i(y_i - \beta x_i - \rho(x_i - z_i\pi)) = 0,$$
$$\sum z_i'(x_i - z_i\pi) = 0.$$

# Standard errors using GMM

But we are short a constraint now that we have introduced $\rho$. We can use the fact that $\hat{x}_i = x_i - \hat{v}_i$ and set as our new conditions, which imply those previously,

$$\sum x_i(y_i - \beta x_i - \rho(x_i - z_i\pi)) = 0,$$
$$\sum \hat{v}_i(y_i - \beta x_i - \rho(x_i - z_i\pi)) = 0,$$
$$\sum z_i'(x_i - z_i\pi) = 0.$$

These conditions are intuitive, because $x$ and $\hat{v}$ are our regressors in our estimating equation.

# Standard errors using GMM

$$\sum x_i(y_i - \beta x_i - \rho v(\pi; x_i, z_i)) = 0,$$
$$\sum \hat{v}_i(y_i - \beta x_i - \rho v(\pi; x_i, z_i)) = 0,$$
$$\sum z_i' v(\pi; x_i, z_i) = 0.$$

We compute GMM-style variance-covariance matrices using moment conditions based on the estimating equation and first stage.

Note the sample variance will thus depend on $G(\hat{\beta}, \hat{\rho}, \hat{\pi})$, the Jacobian with respect to the parameters, evaluated at the optimum. Since $\pi$ appears in the error function of the main equation, we account for dependence between the two equations.

Using the GMM framework lets us easily allow for clustering and HAC VCEs.

# Standard errors using GMM

## Computation

Note: we do not run `gmm` under the hood. We get estimates using the regular two-step procedure. Then GMM standard errors are computed directly in Mata, making the procedure quite fast.

# Postestimation

- `predict`: allowed with statistics `xb`, `xbv`, `e`, `ve`
  - `xb` returns linear prediction without the control function
  - `xbv` returns linear prediction with control function
  - `e` returns the residual not including the control function
  - `ve` returns the residual including the control function
- `margins`: allowed with `xb` and `xbv`
- `estat endogenous`: Translates readily from the corresponding postestimation command for `ivregress`.

# Postestimation example

```
. cfregress rent pcturban (hsngval = faminc i.region, interact(pcturban)), vce(robust)
Control-function linear regression                  Number of obs =      50
                                                    Wald chi2(2)  =   44.83
                                                    Prob > chi2   =  0.0000
                                                    R-squared     =  0.5574
                                                    Root MSE      = 23.2829
Endogenous variable model:
    Linear: hsngval
```

| rent | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| rent | | | | | | |
| hsngval | .0024082 | .0006391 | 3.77 | 0.000 | .0011556 | .0036608 |
| pcturban | .1459889 | .4308807 | 0.34 | 0.735 | -.6985218 | .9904997 |
| _cons | 108.2288 | 17.36071 | 6.23 | 0.000 | 74.20243 | 142.2552 |
| e.rent | | | | | | |
| cf(hsngval) | .0015522 | .0019371 | 0.80 | 0.423 | -.0022444 | .0053488 |
| cf(hsngval)#pcturban | -.0000419 | .0000236 | -1.78 | 0.075 | -.0000881 | 4.26e-06 |

```
Instruments for hsngval: faminc 2.region 3.region 4.region
. estat endogenous
Tests of endogeneity
H0: Variables are exogenous
 ( 1)  [e.rent]cf(hsngval) = 0
 ( 2)  [e.rent]cf(hsngval)#c.pcturban = 0
           chi2(  2) =    7.67
         Prob > chi2 =    0.0216
```

# CF versus IV

## When should I use CF in linear models with endogeneity?

1. When $\hat{\beta}_{CF} = \hat{\beta}_{IV}$, meaning you have no endogenous interactions or fancy first-stage modeling, use plain IV, unless you want the convenient endogeneity test.
2. When you have information about the form of the endogeneity, use CF.
3. When you have endogenous variables entering as interactions, use CF unless you think the IV assumptions are preferable.
4. When you want a nonlinear first stage, use CF.
5. When you have a model that IV commands won't let you run (different instrument sets, exogenous variables excluded from the first stage, etc.), feel free to use CF but make sure you can justify the appropriate assumptions.

# Further reading


- 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.
- Newey, W. K. 1984. A method of moments interpretation of sequential estimators. Economics Letters 14: 201–206.
- Wooldridge, J. M. 2015. Control function methods in applied econometrics. Journal of Human Resources 50: 420–445.