Statistics/Data Analysis

help for **tsls**

Fast and Small 2SLS with FE, IV and Clustered SE

Description

tsls *depvar* *indepvars* **(***varlist2* **=** *varlist_iv***)** [, **areg** **cluster(clusterid)** **demean** **fe(panelid)** **replace**]

This procedure does two-stage least squares with fixed effects, instumental variables and clustered standard errors. While not covering all the capabilities of **xtivreg2** or **ivregress** it is memory efficient and is many times faster. Coeficients and standard errors are unaffected. It is intended for datasets with hundreds of millions of observations and hundreds of variables and for users with time for a bit of care and preparation.

Options

areg fe(*panelid***)** must also be specified. Use the **areg** instead of the **regress** procedure for the second stage regression, absorbing *panelid* with means calculated on-the-fly. This option is incompatible (and unnecessary) with **demean** and **replace**. See notes below. Standard errors are corrected to match **xtivreg**.

cluster(*clusterid***)** Cluster standard errors by *clusterid*, which may be different from *panelid*.

demean Demean the variables by **fe(***panelid***)** before running the regression. This is incompatible and unneccessary with **areg**. If **replace** is specified the demeaning is done in place and the original data is overwritten. This reduces the memory load and if you have multiple regressions with overlapping variables it is efficient to include all your variables in an initial regression with **demean** and then subsequent regressions with only the **fe(***panelid***)**. The first regression will drop rows with missing data, and subsequent regressions will be from the same subsample. Note that if you add an un-demeaned variable in one of the subsequent regressions, there will be no error message but the result will be wrong.

fe(*panelid***)** Specify the variable identifying panel units. If **demean** is not specified this only affects the degrees of freedom.

replace Used with **demean** to cause variables listed in the regression to be replaced with their own deviations from panel unit means.

Examples

Fixed effects with a storage constraint and clustered errors. This doesn't affect the data.

```
. tsls y1 y2,areg fe(panelid) cluster(clusterid)
```

Fixed effect and instrumental variable but the original data is overwritten.

```
. preserve
. tsls y1 (y2 = z1),demean fe(panelid) replace
```

Add clustered standard errors but use the previously demeaned data

```
. tsls y1 (y2=z1) fe(panelid) cluster(clusterid)
```

Drop the IV procedure, still using demeaned data

```
.  tsls y1 y2,fe(panelid)
```

Check the IV result against **xtivreg2**

```
. restore
. xtivreg2 y1 (y2 = z1) vce(clustervar clusterid) absorb(panelid)
```

Notes

Please note that if any regressions expecting demeaned data refer to variables that are not demeaned the result will be incorrect. Hence the order of commands in the example.

Variables listed in {it varlist2} and {it varlist_iv} must not overlap with any variables listed among {it indepvars}.

<cmd:tsls} will always use less memory than **xtivreg** because **xtivreg** stores the demeaned variables as additional doubles. **tsls** stores the demeaned values as floats, and writes over the in-memory original data. If the original data is integer or byte, the option **areg** demeans each row on-the-fly and uses no extra memory at all for demeaned data. If **replace** is specified, there is additional time for input/output.

Standard errors are corrected for degrees of freedom, IV and clustering but you should compare on a subset of your data to **xtivreg2** to confirm this is done correctly. Coeffients and standard errors have matched to the full printed precision in our tests but it is possible we haven't considered every possible situation.

Standard errors in the second stage regression are obtained from a regression of the predicted errors on the RHS variables, but using the true values of the endogenous variables. We would like to thank Doug Staiger for this suggestion, and Jeffrey Wooldridge for noting that because the 2SLS residuals are always uncorrelated in sample with the first-stage fitted values regressing them on the actual data leads to Equation 5.34 of his textbook. We remain resposible for all errors.

This is beta-level software. Please report problems.

More Information

There is more information at http://www.nber.org/stata/tsls . tsls.ado is by Jacob Robbins, tsls.sthlp is by Daniel Feenberg (feenberg@nber.org).

Reference

Wooldridge, Jeffrey M., {it Econometric Analysis of Cross Section and Panel Data}