Linear models and related using Stata

Enrique Pinzón

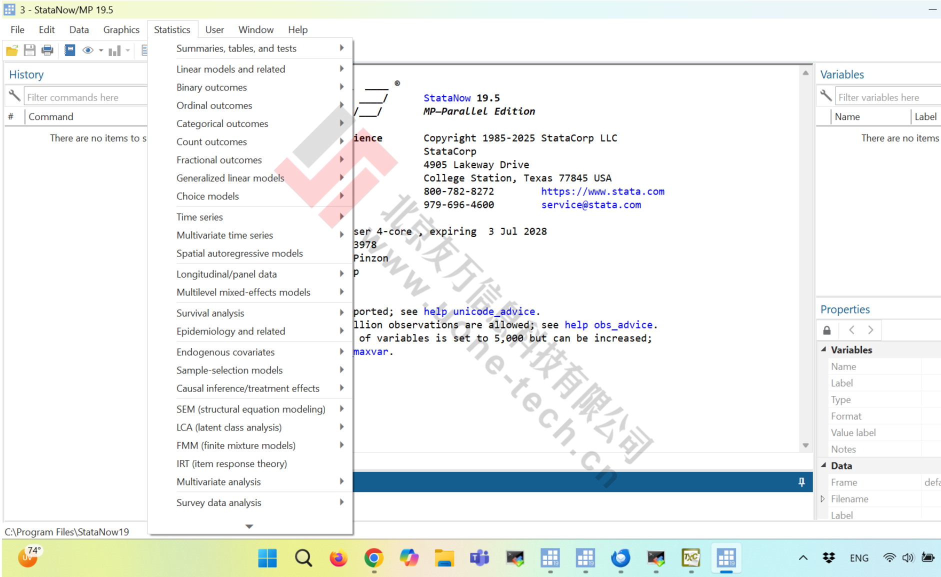
StataCorp LLC

July, 2025

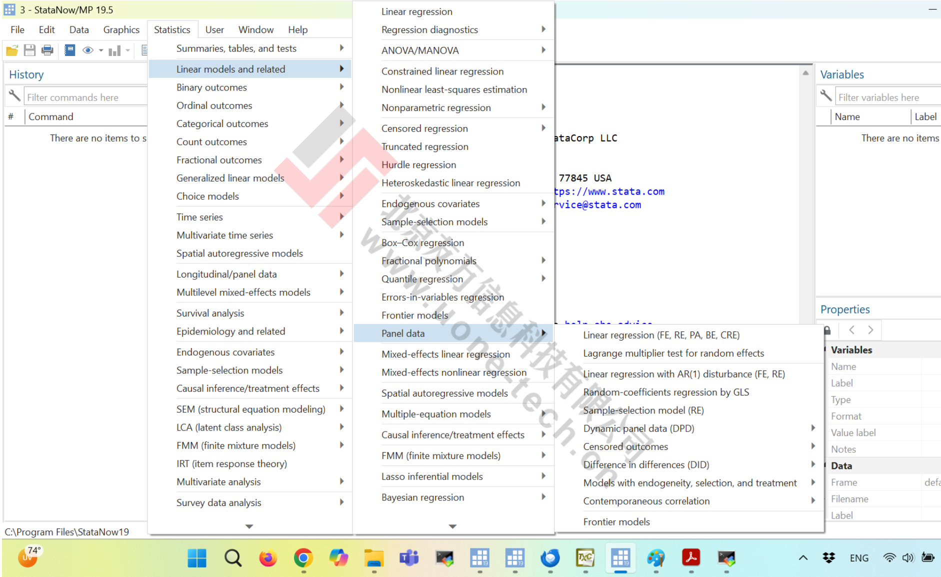
Introduction

- Linear models have been an integral part of Stata since its inception
- Linear models are the most commonly used tools
  - Teaching
  - Build intuition and explore
  - Punching bag
- We are still developing linear regression extension and refinements
- Most current developments that we have incorporated in Stata
- Along the way I will show some simulation and theoretical results

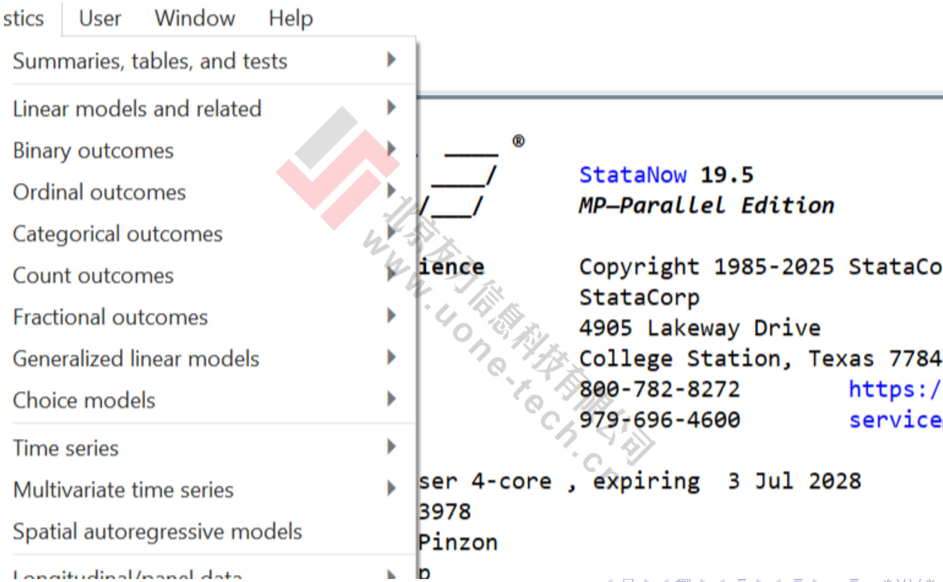
Stata Menu


3 - StataNow/MP 19.5
File
Edit
Data
Graphics
Statistics
User
Window
Help
History
Filter commands here
Command
There are no items to s
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Fractional outcomes
Generalized linear models
Choice models
Time series
Multivariate time series
Spatial autoregressive models
Longitudinal/panel data
Multilevel mixed-effects models
Survival analysis
Epidemiology and related
Endogenous covariates
Sample-selection models
Causal inference/treatment effects
SEM (structural equation modeling)
LCA (latent class analysis)
FMM (finite mixture models)
IRT (item response theory)
Multivariate analysis
Survey data analysis
StataNow 19.5
MP–Parallel Edition
Copyright 1985-2025 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-782-8272 https://www.stata.com
979-696-4600 service@stata.com
ser 4-core , expiring 3 Jul 2028
ported; see help unicode_advice.
llion observations are allowed; see help obs_advice.
of variables is set to 5,000 but can be increased;
maxvar.
Variables
Filter variables here
Name
Label
There are no items
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame
Filename
Label
C:\Program Files\StataNow19
北京友万信息科技有限公司
www.uone-tech.cn
ENG

Linear models and related



Linear models and related



Game plan

- Estimating β of interest
- Inference for $\widehat{\beta}$
- Simulation exercises

Estimating β of interest

High dimensional “fixed effects"

```
. ssc hot
Top 10 packages at SSC
        Jun 2025
 Rank   # hits     Package       Author(s)
 -----------------------------------------------------------------------
    1  166246.7     estout        Ben Jann
    2  116502.7     asdoc         Attaullah Shah
    3  107891.5     outreg2       Roy Wada
    4  100261.0     reghdfe       Sergio Correia
    5  92930.7     winsor2       Yujun Lian
    6  59048.5     ftools        Sergio Correia
    7  54003.7     sum2docx      Chuntao Li, Yuan Xue
    8  45016.8     reg2docx      Chuntao Li, Yuan Xue
    9  40756.0     coefplot      Ben Jann
   10  39717.2     ivreg210      Mark E Schaffer, Steven Stillman,
                                   Christopher F Baum
 -----------------------------------------------------------------------
 (Click on package name for description)
```

A toy model

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{1,502} \gamma_k a_{ik} + \sum_{k=1}^{1,124} \theta_k b_{ik} + \varepsilon_i$$

- We are interested in the marginal effect of x on y
- The categorical variables a and b serve as controls
- The 1,501 parameters of a and 1,123 parameters for b are not relevant.

High dimensional “fixed effects"

```
. timer clear
. timer on 1
. quietly areg y x, absorb(a b)
. timer off 1
. timer on 2
. quietly regress y x i.a i.b
. timer off 2
. timer list
   1:      0.12 /        1 =       0.1240
   2:     75.78 /        1 =      75.7810
```

Fixed effects vs. categorical variables

$$y_{it} = \beta_0 + \beta_1 \mathbf{x}_{it} + \gamma \mathbf{a}_{it} + \theta \mathbf{b}_{it} + \alpha_i + \varepsilon_{it}$$

- α_i is a time invariant unobservable
- α_i is a fixed effect
- $\mathbf{a}_{it}$ and $\mathbf{b}_{it}$ are categorical variables
- In panel data we absorb α_i

syntax areg

- `areg ..., absorb(varlist)`
- `areg` treats data as a cross-section
- Any variable that is absorbed is considered to be a regressor
- In other words there are no incidental parameters or fixed effects
- There is no concept of panel data

syntax areg

- `areg ..., absorb(`*`varlist`*`)`
- `areg` treats data as a cross-section
- Any variable that is absorbed is considered to be a regressor
- In other words there are no incidental parameters or fixed effects
- There is no concept of panel data

syntax xtreg, fe

`xtreg ..., fe absorb(varlist)`

- Any variable that is absorbed in `absorb()` is considered to be a regressor
- The time invariant heterogeneity (fixed effect) is not estimable
- `xtreg` treats data as panel data
  - When errors are not assumed to be i.i.d the “fixed-effects" are not accounted in d.f.a
  - Weights are constant within panel
  - Panels should be nested within cluster

Generate a panel data model

```
clear
set seed 111
set obs 10000
gen id = _n
generate ai = rnormal(0, 4)
expand 10
bysort id: generate time = _n + 2015
generate a  = runiformint(1,1000)
generate b  = runiformint(1,1000)
generate x  = rnormal()
generate e  = rnormal(0,6)
generate y = 10 - 10*x + a/3000 - b/2000 + e + ai
```

xtreg vs. areg I

```
. xtset id time
Panel variable: id (strongly balanced)
 Time variable: time, 2016 to 2025
         Delta: 1 unit
. quietly xtreg y x, absorb(a b) fe
. estimates store fe_iid
. quietly areg  y x, absorb(a b id)
. estimates store areg_idd
. etable, estimates(fe_iid areg_idd) column(command)                    ///
>         cstat(_r_b,nformat(%9.4g)) cstat(_r_se,nformat(%9.4g))
```

	xtreg	areg
x	-10.02	-10.02
	(.02024)	(.02024)
Intercept	9.898	9.898
	(.01893)	(.01893)
Number of observations	100000	100000

xtreg vs. areg II

$$y_{it} = \beta_0 + \beta_1 \mathbf{x}_{it} + \gamma \mathbf{a}_{it} + \theta \mathbf{b}_{it} + \alpha_i + \varepsilon_{it}$$

- Error term is $\alpha_i + \varepsilon_{it}$
- There is intra panel correlation to be accounted for
- Absorbing α_i does not eliminate all correlation

xtreg vs areg II

```
. quietly xtreg y x, absorb(a b) fe vce(robust)
. estimates store fe_robust
. quietly areg  y x, absorb(a b id) vce(robust)
. estimates store areg_robust
. etable, estimates(fe_robust areg_robust) column(command)        ///
>         cstat(_r_b,nformat(%9.4g)) cstat(_r_se,nformat(%9.4g))
------------------------------------------------------------------
                                     xtreg        areg
------------------------------------------------------------------
x                                   -10.02      -10.02
                                  (.01999)    (.02012)
Intercept                            9.898       9.898
                                (5.05e-06)    (.01893)
Number of observations              100000      100000
------------------------------------------------------------------
```

More linear models

- `didregress`
- `xtdidregress`
- `xthdidregress twfe`
- `xtreg, cre`
- All of this are using `_regress` except for `xtreg, cre`

More linear models

- `didregress`
- `xtdidregress`
- `xthdidregress twfe`
- `xtreg, cre`
- All of this are using `_regress` except for `xtreg, cre`

didregress

```
didregress (y xvars) (treatment), group(group) time(t)
```

- `areg y xvars i.t treatment, absorb(group) ///`
  `vce(cluster group)`
- `treatment` indicator for being in a treated group in the period after intervention
- `treatment` is an interaction of a treated group and post period
- All individuals in a group should have the same treatment status at the same time
- There should be a control group (never treated)
- Standard errors are cluster robust by default

didregress

```
didregress (y xvars) (treatment), group(group) time(t)
```

- `areg y xvars i.t treatment, absorb(group) /// vce(cluster group)`
- `treatment` indicator for being in a treated group in the period after intervention
- `treatment` is an interaction of a treated group and post period
- All individuals in a group should have the same treatment status at the same time
- There should be a control group (never treated)
- Standard errors are cluster robust by default

xthdidregress twfe

```
xthdidregress twfe (y xvars) (d), group(group) time(t)
```

- regress `y` on `cohort, time, xvars` and interactions and their panel level means
- a regression with panel level means of regressors is a Mundlak regression (CRE)
- cohort variable is created (defines first time a group is treated)
- compute contrast of treatment over cohorts at a given time (`margins`)
- variance covariance matrix is `vce(unconditional)`, takes into account variation in covariates

AKC data

```
. use akc
(Fictional dog breed and AKC registration data)
. list in 1/15, sepby(breed) noobs abbreviate(20)
```

year	breed	movie	best	registered
2031	Affenpinscher	0	0	1653
2032	Affenpinscher	0	0	1340
2033	Affenpinscher	0	0	1180
2034	Affenpinscher	0	0	1602
2035	Affenpinscher	0	0	934
2036	Affenpinscher	0	0	497
2037	Affenpinscher	0	0	1395
2038	Affenpinscher	0	0	1656
2039	Affenpinscher	0	0	1663
2040	Affenpinscher	0	0	1166
2031	Afghan Hound	0	0	1341
2032	Afghan Hound	0	0	1398
2033	Afghan Hound	0	0	1544
2034	Afghan Hound	0	0	791
2035	Afghan Hound	0	0	531

xtreg, cre (etable)

```
. quietly use akc, clear
. quietly xtset breed year
. bysort breed: egen mmovie = mean(movie)
. quietly xtreg registered movie, fe vce(robust)
. estimates store fe
. quietly regress registered movie mmovie, vce(cluster breed)
. estimates store mundlak
. quietly xtreg registered movie, cre vce(robust)
. estimates store cre
. etable, estimates(fe mundlak cre)
```

	registered	registered	registered
Was a movie protagonist	2185.141	2185.141	
	(69.375)	(69.400)	
mmovie		-230.653	
		(70.930)	
Intercept	1011.490	1028.339	
	(5.068)	(13.168)	
Was a movie protagonist			2185.141
			(69.400)
Intercept			1028.339
			(13.168)
Was a movie protagonist			-230.653
			(70.930)
Number of observations	1410	1410	1410

xtreg, cre

```
. quietly use akc, clear
. quietly xtset breed year
. bysort breed: egen mmovie = mean(movie)
. quietly xtreg registered movie, fe vce(robust)
. estimates store fe
. quietly regress registered movie mmovie, vce(cluster breed)
. estimates store mundlak
. quietly xtreg registered movie, cre vce(robust)
. estimates store cre
. etable, estimates(fe mundlak cre) column(estimates)          ///
>        cstat(_r_b,nformat(%9.4g)) cstat(_r_se,nformat(%9.4g)) ///
>                eqrecode(registered = xb registered = xb xit_vars =xb)
```

	fe	mundlak	cre
Was a movie protagonist	2185	2185	2185
	(69.38)	(69.4)	(69.4)
mmovie		-230.7	
		(70.93)	
Intercept	1011	1028	1028
	(5.068)	(13.17)	(13.17)
Was a movie protagonist			-230.7
			(70.93)
Number of observations	1410	1410	1410

Inference for $\widehat{\beta}$

What is new in inference

- Wild cluster bootstrap
- HC2 and HC3 for clusters
- HC2 and HC3 for clusters with degrees of freedom adjustment
- HC3 with Hansen degrees of freedom adjustment and rescaling
- Multiway clustering

Why do we care

- Bertrand, Duflo, and Mulainathan (2004)
- Cameron and Miller (2015)
- Imbens and Kolesar (2016)
- McKinnon, Nielsen, and Webb (2023)
- Hansen (2024)

Wild cluster bootstrap

```
wildbootstrap estimator y xvars [if] [in] [weights], opts
```

- *estimator*: areg, regress, xtreg
- Computes confidence intervals and p-values
- Does not compute standard errors

Wild cluster bootstrap

1. regress y on $x_1 \ldots x_k$ obtain t-statistic
2. impose constraint $\beta_j = 0$
3. run regression imposing constraint and obtain $\tilde{\beta}$ and $\tilde{\varepsilon}$
4. generate $\tilde{y} = x\tilde{\beta} + \tilde{\varepsilon}\nu$
5. run regression using $\tilde{y}$ and original covariates

wilbootstrap example

```
. wildbootstrap xtreg registered movie, rseed(111)
Panel variable: breed (strongly balanced)
 Time variable: year, 2031 to 2040
         Delta: 1 unit
Performing 1,000 replications for p-value for movie = 0 ...
Computing confidence interval for movie
  Lower bound: .........10.........20.. done (22)
  Upper bound: .........10.........20.. done (22)
Wild cluster bootstrap                              Number of obs      = 1,410
Fixed-effects linear regression                     Number of clusters =   141
                                                    Cluster size:
Cluster variable: breed                                          min =    10
Error weight: Rademacher                                         avg =  10.0
                                                                 max =    10
--------------------------------------------------------------------------------
           registered |   Estimate       t   p-value    [95% conf. interval]
----------------------+---------------------------------------------------------
constraint            |
            movie = 0 |   2185.141   31.50   0.000     2029.609    2322.348
--------------------------------------------------------------------------------
```

HC2

$$\widehat{\mathbf{V}}_{hc2cluster} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{M}_g^{+1/2} \widehat{\epsilon}_g \widehat{\epsilon}_g \mathbf{M}_g^{+1/2} \mathbf{X}_g \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- $\mathbf{M}_g^{+1/2} = \left(\mathbf{I}_{n_g} - \mathbf{X}_g \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}_g' \right)^{+1/2}$ is an eliminator matrix and we are using a Moore-Penrose inverse
- Imbens and Kolesar found good properties for cluster case
- It is a generalization of HC2 with a degrees of freedom adjustment of Bell and McCaffrey

HC2 syntax

- *estimator* `..., vce(hc2 [clustervar], [dfadjust]`
  - `areg`
  - `xtreg`
  - `regress`

HC2

```
. xtdidregress (y x1 x2 i.a i.b) (digt), group(G) time(t) vce(hc2)
Computing degrees of freedom ...
Treatment and time information
Time variable: t
Control:       digt = 0
Treatment:     digt = 1
-----------------------------------
             |   Control  Treatment
-------------+---------------------
Group        |
           G |        46          4
-------------+---------------------
Time         |
     Minimum |         1          2
     Maximum |         1          2
-----------------------------------

Difference-in-differences regression                 Number of obs   = 30,000
                                                     No. of clusters =     50
Data type: Longitudinal
                                   (Std. err. adjusted for 50 clusters in id)
------------------------------------------------------------------------------
             |             Robust HC2
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
        digt |
   (1 vs 0)  |  -.0653023   .1189582    -0.55   0.610     -.385554    .2549495
------------------------------------------------------------------------------
Note: ATET estimate adjusted for covariates, panel effects, and time effects.
. estimates store hc2
```

HC3

$$\widehat{\mathbf{V}}_{hc2cluster} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{M}_g^{+} \widehat{\epsilon}_g \widehat{\epsilon}_g \mathbf{M}_g^{+} \mathbf{X}_g \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- $\mathbf{M}_g^{+} = \left(\mathbf{I}_{n_g} - \mathbf{X}_g \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}_g'\right)^{+}$ is an eliminator matrix and we are using a Moore-Penrose inverse
- Suggested as a possibility by Cameron and Miller (2015)
- It is a generalization of HC2 with a degrees of freedom adjustment of Bell and McCaffrey

HC3 syntax

- *estimator* ..., vce(hc3 [clustervar], [dfadjust hansen]
  - areg
  - xtreg
  - regress

HC3

```
. xtdidregress (y x1 x2 i.a i.b) (digt), group(G) time(t) vce(hc3)
Computing degrees of freedom ...
Treatment and time information
Time variable: t
Control:       digt = 0
Treatment:     digt = 1
-----------------------------------
             |   Control  Treatment
-------------+---------------------
Group        |
           G |        46          4
-------------+---------------------
Time         |
     Minimum |         1          2
     Maximum |         1          2
-----------------------------------
Difference-in-differences regression                 Number of obs   = 30,000
                                                     No. of clusters =     50
Data type: Longitudinal
                                    (Std. err. adjusted for 50 clusters in id)
------------------------------------------------------------------------------
             |             Robust HC3
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
        digt |
   (1 vs 0)  |  -.0653023   .1344837    -0.49   0.653    -.4400695     .309465
------------------------------------------------------------------------------
Note: ATET estimate adjusted for covariates, panel effects, and time effects.
. estimates store hc3
```

HC3 with Hansen D.F. Adjustment Example

```
. xtdidregress (y x1 x2 i.a i.b) (digt), group(G) time(t) vce(hc3, hansen)
Computing degrees of freedom ...
Treatment and time information
Time variable: t
Control:       digt = 0
Treatment:     digt = 1
-----------------------------------
             |   Control  Treatment
-------------+---------------------
Group        |
           G |        46          4
-------------+---------------------
Time         |
     Minimum |         1          2
     Maximum |         1          2
-----------------------------------

Difference-in-differences regression                 Number of obs   = 30,000
                                                      No. of clusters =     50
Data type: Longitudinal
                                     (Std. err. adjusted for 50 clusters in id)
------------------------------------------------------------------------------
             |              Robust HC3
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
        digt |
   (1 vs 0)  |  -.0653023   .1344837    -0.49   0.611    -.3955562    .2649517
------------------------------------------------------------------------------
Note: ATET estimate adjusted for covariates, panel effects, and time effects.
Note: p-values and confidence intervals computed using Hansen adjustment.
. estimates store hchansen
```

Simulation exercises

DID DGP

$$
\begin{aligned}
y_{igt}(0) &= \epsilon_{igt} + u_g + h_{ig}\nu_g \\
\epsilon_{igt} &\sim N(0,1) \\
u_g &\sim N(0,1) \\
\nu_g &\sim N(0,1) \\
h_{ig} &= 1 - 2\left(\mathbb{1}\left\{U(0,1) > .5\right\}\right) \\
y_{igt}(1) &= y_{igt}(0) + \varepsilon_{ig} \\
\varepsilon_{ig} &\sim N(0,1) \\
y_{igt} &= d_{igt}\left(y_{igt}(1) - y_{igt}(0)\right) + \mathbf{x}\beta + \alpha_i + \epsilon_{igt}
\end{aligned}
$$

Simulation parameters

- Number of clusters: 10 (small) or 50 (large)
- Cluster sizes: homogeneous (uniform distribution) or heterogeneous (truncated beta)
- Panel of 3000 individuals and 10 time periods (fixed across designs)
- 4 treated clusters in all designs
- Treatment happens for all in period 2

Many homogeneous clusters

Estimator	Rejection rate
HC1	.821
HC2 DF	.943
HC3 DF	.966
HC3 Hansen	.948
Wildboot R	.883
Wildboot W	.898

Few homogeneous clusters

Estimator	Rejection rate
HC1	.886
HC2 DF	.923
HC3 DF	.954
HC3 Hansen	.927
Wildboot R	.909
Wildboot W	.900

Many heterogeneous clusters

Estimator	Rejection rate
HC1	.871
HC2 DF	.934
HC3 DF	.956
HC3 Hansen	.938
Wildboot R	.919
Wildboot W	.913

Few heterogeneous clusters

Estimator	Rejection rate
HC1	.890
HC2 DF	.913
HC3 DF	.938
HC3 Hansen	.918
Wildboot R	.906
Wildboot W	.900

Takeaways

- Default cluster robust standard errors tend to under-reject
- HC2 and HC3 standard errors with degrees of freedom correction perform well
- HC3 with Hansed D.F. correction has a similar behavior to other D.F. adjustment methods

Conclusion

- Linear models and related have new and important features in the last two releases
- We are still working and improving to add more features both in estimation and inference
- We have added the `absorb()` option to `areg` and `xtreg, fe` with important interpretation differences
- HC2 and HC3 estimator with degrees of freedom seem to be a good alternative when there are few and heterogeneous clusters

Conclusion

- Linear models and related have new and important features in the last two releases
- We are still working and improving to add more features both in estimation and inference
- We have added the `absorb()` option to `areg` and `xtreg, fe` with important interpretation differences
- HC2 and HC3 estimator with degrees of freedom seem to be a good alternative when there are few and heterogeneous clusters