Stata 19—
see the new features in action

Meghan Cain, StataCorp LLC

23 September 2025

New features in Stata 19

https://www.stata.com/new-in-stata/

- Machine learning via H2O: Ensemble decision trees
- Conditional average treatment effects (CATE)
- High-dimensional fixed effects (HDFE)
- Bayesian variable selection for linear model
- Interval-censored multiple-event Cox model
- Meta-analysis for correlations
- Bayesian bootstrap
- Control-function linear and probit models
- Bayesian quantile regression
- Inference robust to weak instruments
- Mundlak specification test
- Correlated random-effects (CRE) model
- Panel-data vector autoregressive (VAR) model
- SVAR models via instrumental variables
- Instrumental-variables local-projection IRFs
- Latent class model-comparison statistics
- Bayesian asymmetric Laplace model
- Do-file Editor: Autocompletion, templates, and more
- Graphics: Bar graph CIs, heat maps, and more
- Tables: Easier tabulations, exporting, and more
- Multiple datasets: Modify a set of frames
- Stata in French
- More

New features in Stata 19

Machine learning

- **Machine learning via H2O: Ensemble decision trees**

Causal inference

- Conditional average treatment effects (CATE)
- Control-function linear and probit models
- Inference robust to weak instruments

Time series

- SVAR models via instrumental variables
- Instrumental-variables local-projection IRFs

Panel data

- Panel-data vector autoregressive (VAR) model
- **High-dimensional fixed effects (HDFE)**
- Correlated random-effects (CRE) model
- Mundlak specification test

Bayesian

- Bayesian variable selection for linear model
- Bayesian bootstrap
- Bayesian quantile regression
- Bayesian asymmetric Laplace model

Other stats: Survival/Meta/SEM

- **Interval-censored multiple-event Cox model**
- Meta-analysis for correlations
- Latent class model-comparison statistics

Workflow

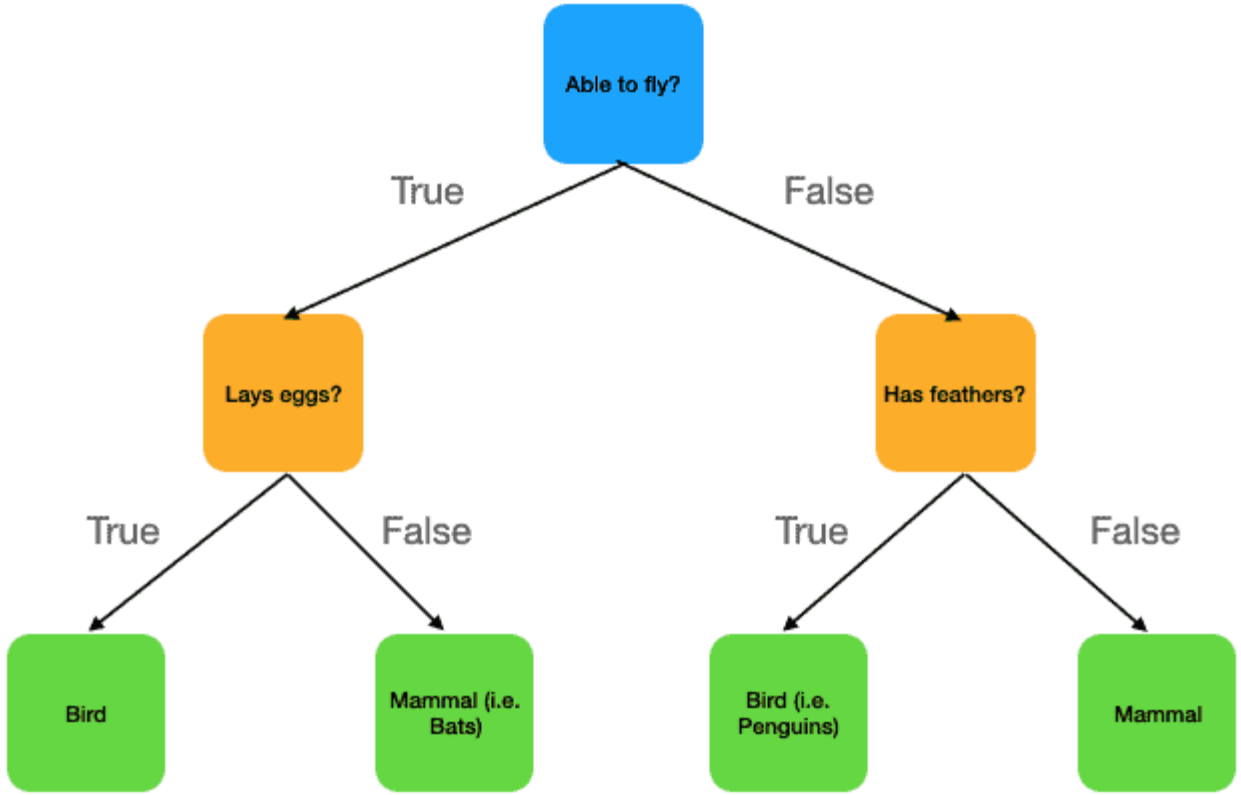
- **Do-file Editor: Autocompletion, templates, and more**
- Multiple datasets: Modify a set of frames

Graphics and reporting

- Graphics: Bar graph CIs, heat maps, and more
- Tables: Easier tabulations, exporting, and more

Machine learning via H2O: Ensemble decision trees

- Gradient boosting machine (GBM) and random forest
- Regression, binary classification, and multiclass classification
- Hyperparameter tuning, model performance, and prediction
- Cross-validation (CV) and grid-search
- Prediction explainability
- Variable importance



https://insidelearningmachines.com/interpret_decision_trees/

Conditional average treatment effects (CATE)

```
. cate po (assets $covars) (e401k), group(incomecat)

Conditional average treatment effects     Number of observations       = 9,913
Estimator:       Partialing out           Number of folds in cross-fit =    10
Outcome model:   Linear lasso             Number of outcome controls   =    17
Treatment model: Logit lasso              Number of treatment controls =    17
CATE model:      Random forest            Number of CATE variables     =    17
```

assets	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
GATE						
incomecat						
0	4009.913	997.1495	4.02	0.000	2055.536	5964.29
1	1396.478	1656.966	0.84	0.399	-1851.116	4644.072
2	5089.059	1346.355	3.78	0.000	2450.25	7727.867
3	8645.766	2294.041	3.77	0.000	4149.528	13142
4	20440.61	4713.003	4.34	0.000	11203.3	29677.93
ATE						
e401k						
(Eligible vs Not eligible)	7916.781	1151.562	6.87	0.000	5659.761	10173.8
POmean						
e401k						
Not eligible	14027.23	831.7175	16.87	0.000	12397.1	15657.37

Control-function linear and probit models

```
. cfregress lndrug age lninc (ins = married work, probit interact(ins)), mainonly(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
```

lndrug	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
lndrug						
1.ins	-.8598836	.3483648	-2.47	0.014	-1.542666	-.1771011
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	.2702095	.2585099	1.05	0.296	-.2364605	.7768796

```
Instruments for 1.ins: married work
```

Inference robust to weak instruments

```
. ivregress 2sls lndrug age lninc (ins = married work), vce(robust)


. estat weakrobust

Test robust to weak instruments
Model VCE: Robust

 ( 1)  ins = 0

Cond. likelihood-ratio (CLR) test =  10.50
                      Prob > CLR = 0.0012

Notes: CLR test reported by default because
       model is overidentified.
       p-value computed by simulation
       (25,000 replications).
```

SVAR models via instrumental variables

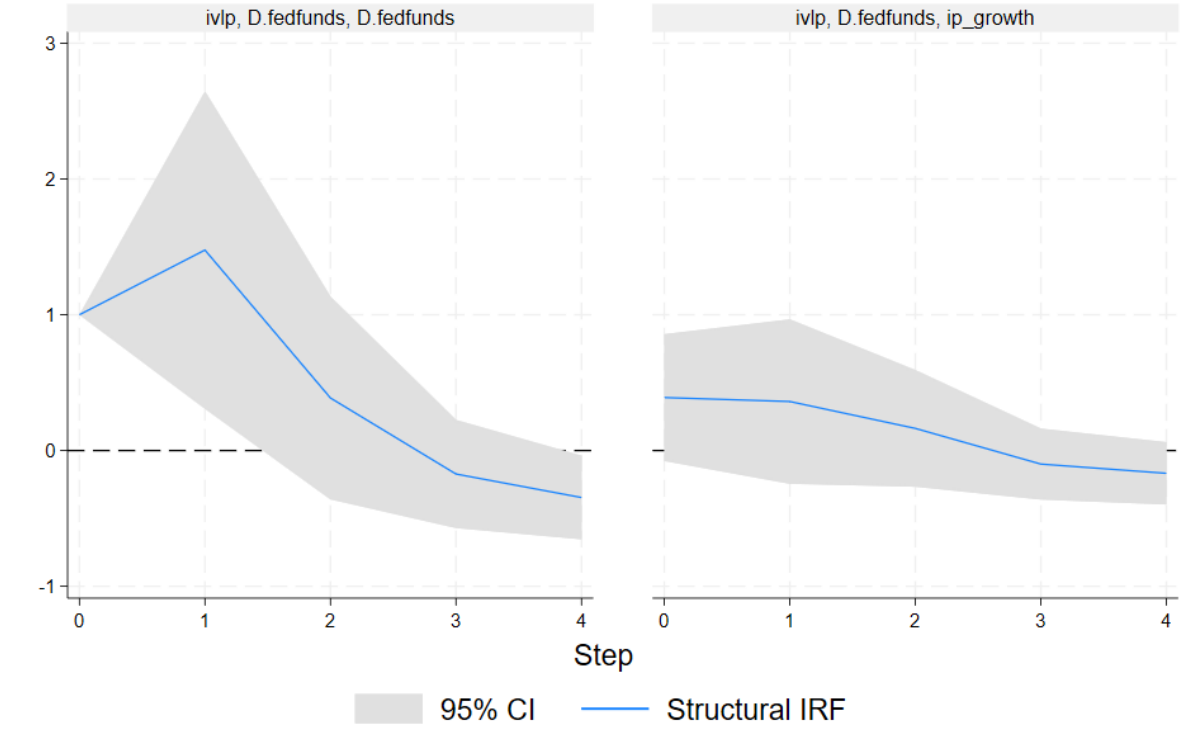
```
. ivsvar gmm ip_growth fedfunds (inflation = oil_inst)

Final GMM criterion = 4.88e-32

note: model is exactly identified.

Instrumental-variables SVAR                              Number of obs = 783
VAR sample: 1954m10 thru 2019m12
GMM sample: 1954m10 thru 2019m12

 ( 1)  [e.inflation]inflation = 1
```

Effect	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
e.inflation						
ip_growth	-.31198	.4330713	-0.72	0.471	-1.160784	.5368241
fedfunds	.0046142	.271441	0.02	0.986	-.5274004	.5366288
inflation	1	(constrained)				

```
Note: Underlying VAR fit with 2 lags.
Dependent variables: ip_growth fedfunds inflation
Instrumented shock: inflation
Instrument: oil_inst
```

IV local-projection IRFs

```
. ivlpirf ip_growth, endogenous(d.fedfunds = money_inst)

. irf set ivlp.irf, replace
. irf create ivlp
. irf graph sirf, yline(0) impulse(D.fedfunds)
```



Graphs by irfname, impulse variable, and response variable

Panel-data vector autoregression (VAR)

```
. xtvar grants revenues expenditures, lags(4)

Panel-data vector autoregression                    Number of obs    = 1,060
Group variable: idcode                              Number of groups =   265
Time variable: year                                 Obs per group:
                                                                min =     4
Number of moment conditions = 198                               avg =   4.0
                                                                max =     4

Fixed-effects transform: FOD
Two-step results
                                        (Std. err. adjusted for 265 clusters in idcode)
------------------------------------------------------------------------------
             |             WC robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
grants       |
      grants |
         L1. |  -.1322887   .1173213    -1.13   0.259    -.3622342    .0976568
         L2. |  -.1067133   .0631209    -1.69   0.091    -.2304281    .0170014
         L3. |    .009167   .0554286     0.17   0.869    -.0994711    .1178051
         L4. |  -.0038156   .0443719    -0.09   0.931    -.0907829    .0831516
             |
    revenues |
         L1. |  -.0205415   .0244862    -0.84   0.402    -.0685336    .0274507
         L2. |  -.0006096    .020736    -0.03   0.977    -.0412515    .0400322
         L3. |  -.0027285   .0189318    -0.14   0.885    -.0398341    .0343771
         L4. |  -.0333568   .0152122    -2.19   0.028    -.0631722   -.0035413
             |
expenditures |
         L1. |  -.0113169   .0243388    -0.46   0.642    -.0590201    .0363863
         L2. |  -.0035009   .0191822    -0.18   0.855    -.0410974    .0340956
         L3. |  -.0053186   .0205777    -0.26   0.796      -.04565    .0350129
         L4. |  -.0262697   .0161428    -1.63   0.104    -.0579091    .0053696
-------------+----------------------------------------------------------------
revenues     |
      grants |
```

High-dimensional fixed effects (HDFE)

Absorb multiple high-dimensional categorical variables:

- Linear models with **areg, absorb()**
- Fixed-effects linear models with **xtreg, fe absorb()**
- Two-stage least-squares regression with **ivregress 2sls, absorb()**

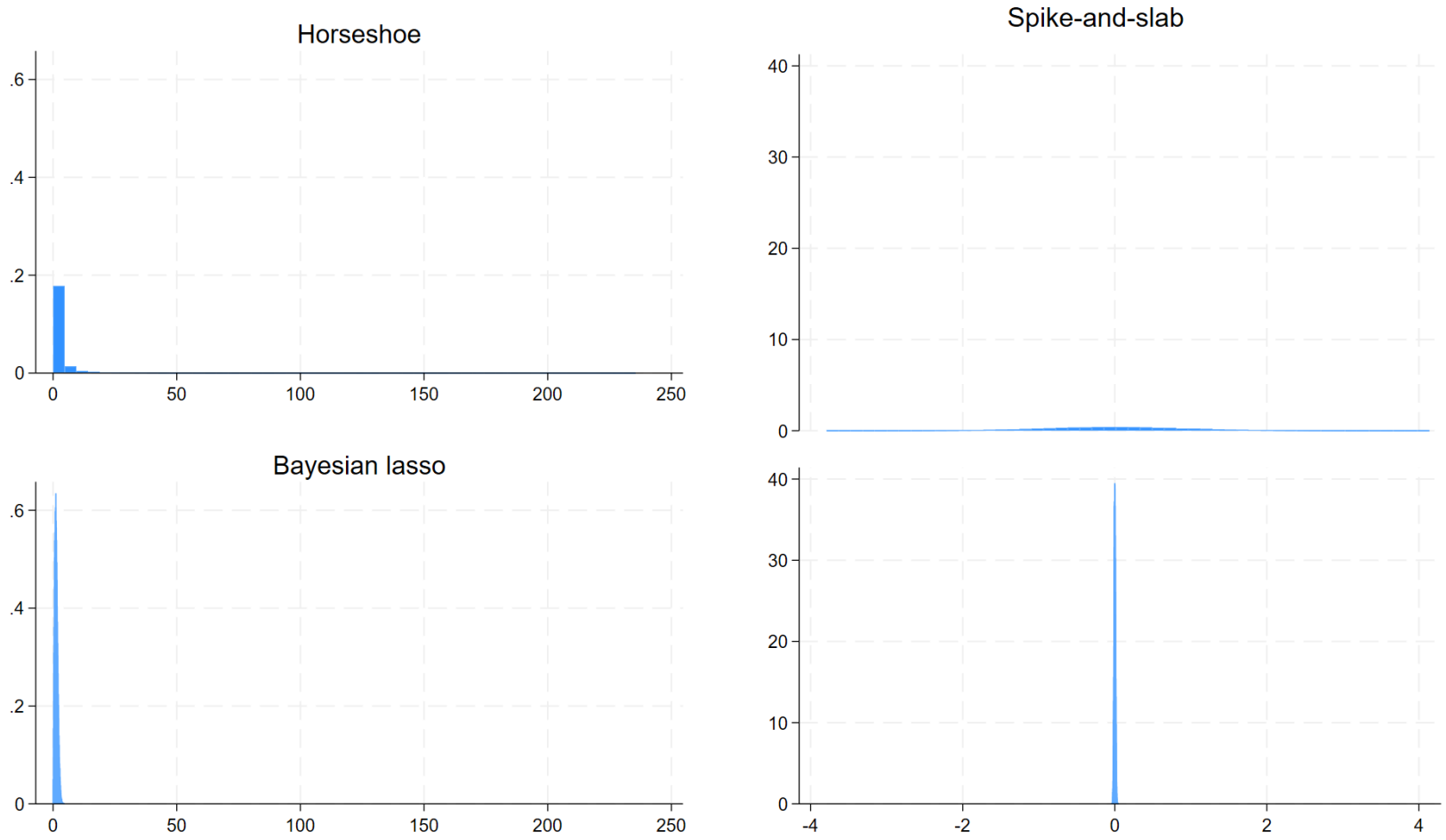
Correlated random-effects (CRE) model

```
. xtreg ln_wage tenure age i.collgrad, cre vce(cluster idcode)
note: 1.collgrad omitted from xt_means because of collinearity.

Correlated random-effects regression            Number of obs     =     28,101
Group variable: idcode                          Number of groups  =      4,699

R-squared:                                      Obs per group:
     Within  = 0.1296                                         min =          1
     Between = 0.3368                                         avg =        6.0
     Overall = 0.2538                                         max =         15

                                                Wald chi2(3)      =    2427.11
corr(xit_vars*b, xt_means*γ) = 0.6144           Prob > chi2       =     0.0000

                                (Std. err. adjusted for 4,699 clusters in idcode)
------------------------------------------------------------------------------
             |               Robust
     ln_wage |  Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
xit_vars     |
      tenure |   .0211313   .0012113    17.45   0.000     .0187573    .0235054
         age |   .0121949   .0007414    16.45   0.000     .0107417     .013648
  1.collgrad |   .3837348   .0133789    28.68   0.000     .3575127    .4099569
       _cons |    1.29133   .0287932    44.85   0.000     1.234896    1.347764
-------------+----------------------------------------------------------------
xt_means     |
      tenure |   .0309138   .0022772    13.58   0.000     .0264506     .035377
         age |  -.0070532   .0012917    -5.46   0.000     -.009585   -.0045215
  1.collgrad |          0  (omitted)
-------------+----------------------------------------------------------------
     sigma_u |  .30438511
     sigma_e |  .29808194
         rho |  .51046112   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Mundlak test (xt_means = 0): chi2(2) = 184.5221          Prob > chi2 = 0.0000
```

Mundlak specification test

```
. estat mundlak

Mundlak specification test
H0: Covariates are uncorrelated with unobserved panel-level effects

    chi2(2) = 184.52
Prob > chi2 = 0.0000

Notes: Fixed effects and correlated random effects are
       consistent under H0 and Ha.
       Random effects are efficient under H0.
```

Bayesian variable selection for linear model



Bayesian variable selection for linear model

```
. bayesselect bwt age-ftv, sslaplace

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  bwt ~ normal(xb_bwt,{sigma2})

Priors:
  {bwt:age ... ftv} ~ mixlaplace(1,.01,1,{gammas})                      (1)
        {bwt:_cons} ~ normal(0,10000)                                   (1)
           {sigma2} ~ jeffreys

Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_bwt.

Bayesian variable selection                     MCMC iterations  =     12,500
Metropolis–Hastings and Gibbs sampling          Burn-in          =      2,500
                                                MCMC sample size =     10,000
Spike-and-slab coefficient prior:               Number of obs    =        189
  Laplace mixture: L(0,.01) and L(0,1)          Acceptance rate  =      .8629
  Beta(1,1) for {theta}                         Efficiency:  min =     .09745
                                                             avg =      .7982
Log marginal-likelihood = -1539.8449                         max =          1

------------------------------------------------------------------------------
             |                                                 Equal-tailed     Inclusion
         bwt |      Mean   Std. dev.     MCSE  [95% cred. interval]     prob.
-------------+----------------------------------------------------------------
         lwt |   18.58596   .9662427  .0165007   16.58558   20.39802      1.00
         age |   2.135033   3.021765  .0967963  -.8348263   10.45639      0.78
       smoke |    .0086326   1.194139  .0119414  -2.652357   2.669007      0.69
          ht |  -.0325025    1.16083  .0118228  -2.670201   2.466308      0.69
         ftv |   .0042764   1.154309  .0115431  -2.535679    2.56472      0.68
          ui |  -.0050708   1.187962   .011994  -2.706212   2.619716      0.68
```

Bayesian bootstrap

```
. bayesboot, priorpowers(priorvar) rseed(111): regress ln_wage tenure
(running regress on estimation sample)

Bayesian bootstrap replications (50): .........10.........20.........30.........40.........50 done

Bayesian bootstrap
Observation prior: priorvar

Linear regression                                   Number of obs =   28,101
                                                    Replications  =       50
                                                    Wald chi2(1)  = 20313.37
                                                    Prob > chi2   =   0.0000
                                                    R-squared     =   0.1373
                                                    Adj R-squared =   0.1373
                                                    Root MSE      =   0.4438
```

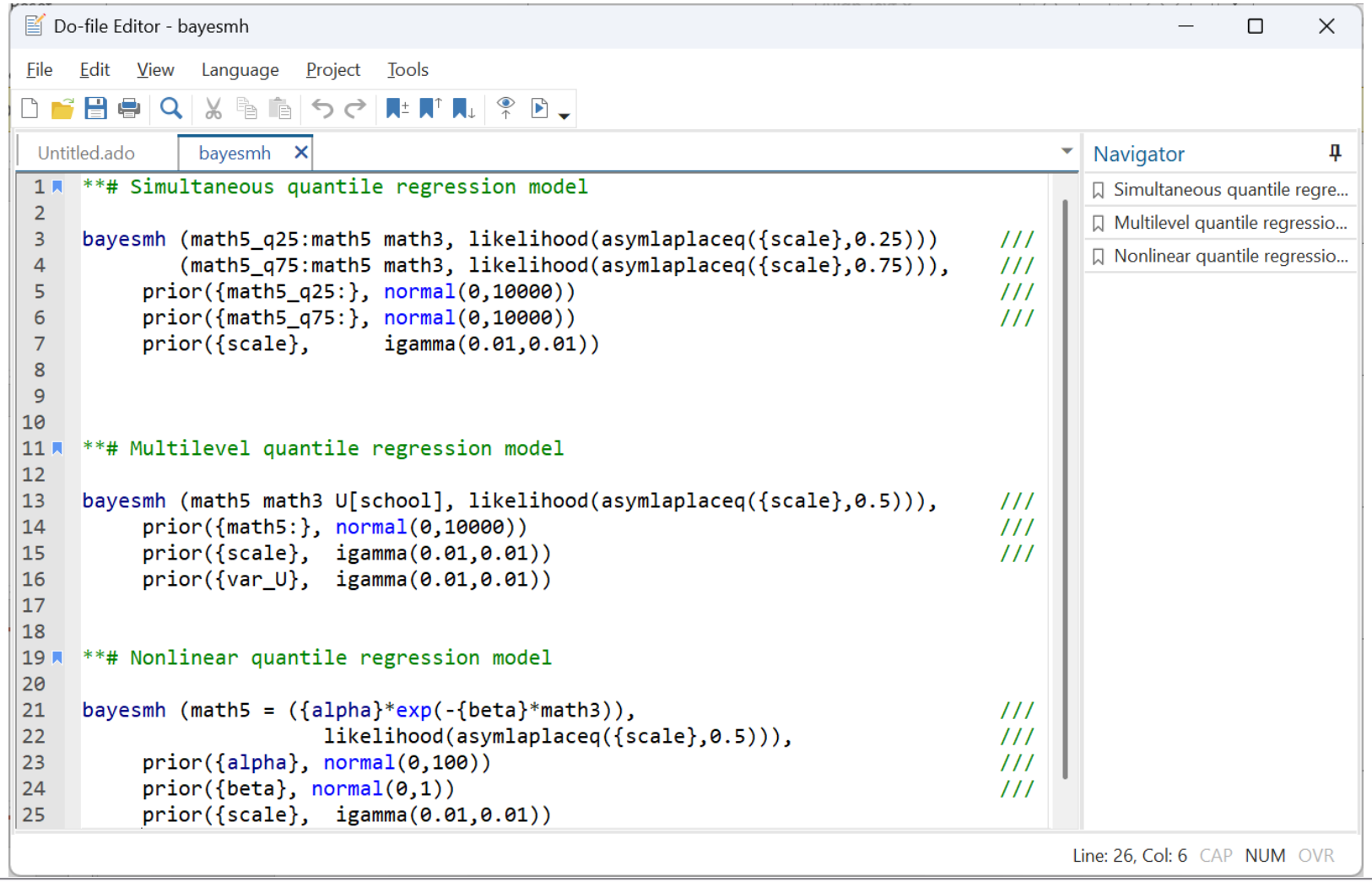
ln_wage	Observed coefficient	Bayesian bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
tenure	.0471994	.0003312	142.52	0.000	.0465503	.0478485
_cons	1.529661	.0017074	895.88	0.000	1.526315	1.533008

Bayesian quantile regression

```
. bayes, rseed(19): qreg ln_wage tenure

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  ln_wage ~ asymlaplaceq(xb_ln_wage_q50,{sigma},.5)

Priors:
  {ln_wage_q50:tenure _cons} ~ normal(0,10000)                          (1)
                      {sigma} ~ igamma(0.01,0.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_ln_wage_q50.

Bayesian quantile regression                  MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling      Burn-in          =      2,500
                                              MCMC sample size =     10,000
Quantile = .5                                 Number of obs    =     28,101
                                              Acceptance rate  =      .3345
                                              Efficiency:  min =      .1158
                                                           avg =      .1567
Log marginal-likelihood = -16568.705                       max =      .2378
```

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]	
ln_wage_q50						
tenure	.052112	.0006169	.000018	.0521365	.050841	.0532015
_cons	1.488457	.0031724	.000093	1.488359	1.482425	1.494786
sigma	.165675	.0010033	.000021	.1656614	.1636896	.1676923

Bayesian asymmetric Laplace model



Interval-censored multiple-event Cox model

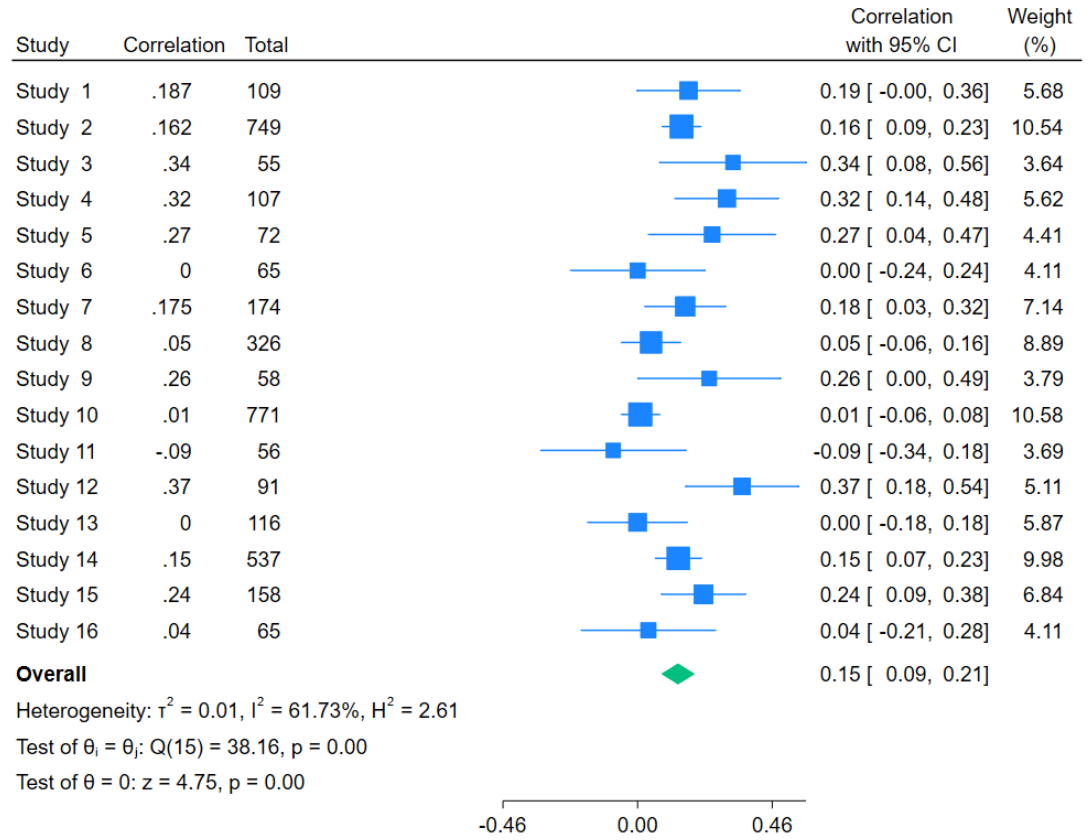
- Use **stmgintcox** to analyze interval-censored multiple-event data and account for possible correlation between event times across the different events.

- **stmgintcox** implements a novel marginal proportional hazards Cox model (Xu, Zeng, and Lin 2023), which is not available in other general-purpose statistical software, commercial or not.

- Left-censoring, right-censoring, interval-censoring
- Support for time-varying covariates.
- Robust and cluster–robust standard errors

Meta-analysis for correlations

```
. meta esize rho n, fisherz
. meta forestplot, correlation
```



Latent class model-comparison statistics

```
. gsem (accident play insurance stock <- ), logit lclass(C 1)
. estimates store lc1
. gsem (accident play insurance stock <- ), logit lclass(C 2)
. estimates store lc2
. gsem (accident play insurance stock <- ), logit lclass(C 3)
. estimates store lc3
. lcstats lc1 lc2 lc3

Latent class statistics
```

	Classes	N	ll	Rank	Entropy	df	LMR	P>LMR
lc1	1	216	-543.65	4				
lc2	2	216	-504.47	9	0.7193	5	75.55	<0.001
lc3	3	216	-503.30	14	0.6110	5	2.25	0.687

```
LMR is the Lo–Mendell–Rubin-adjusted likelihood-ratio test statistic.
Likelihood-ratio tests compare the given model versus the same model
with one less latent class.
```

Latent class model-comparison statistics

```
. gsem (accident play insurance stock <- ), logit lclass(C 1)
. estimates store lc1
. gsem (accident play insurance stock <- ), logit lclass(C 2)
. estimates store lc2
. gsem (accident play insurance stock <- ), logit lclass(C 3)
. estimates store lc3
. lcstats lc1 lc2 lc3

Latent class statistics
```

	Classes	N	ll	Rank	Entropy	df	LMR	P>LMR
lc1	1	216	-543.65	4				
lc2	2	216	-504.47	9	0.7193	5	75.55	<0.001
lc3	3	216	-503.30	14	0.6110	5	2.25	0.687

```
LMR is the Lo–Mendell–Rubin-adjusted likelihood-ratio test statistic.
Likelihood-ratio tests compare the given model versus the same model
with one less latent class.

. collect ...
```

Classes	BIC	LMR (P>LMR)	Class marginal probabilities (SE)		
1	1,108.80		1.00 (0.00)		
2	1,057.31	75.55 (<0.001)	0.72 (0.06)	0.28 (0.06)	
3	1,081.86	2.25 (0.687)	0.16 (15.31)	0.63 (11.94)	0.21 (3.37)

New in Do-file Editor

- Customization of autocompletion
- File templates
- Customization of highlighting of matches
- Bracket highlighting
- Code-folding enhancements
- Temporary bookmarks
- Show whitespace in selection
- Navigator panel

Multiple datasets: Modify a set of frames

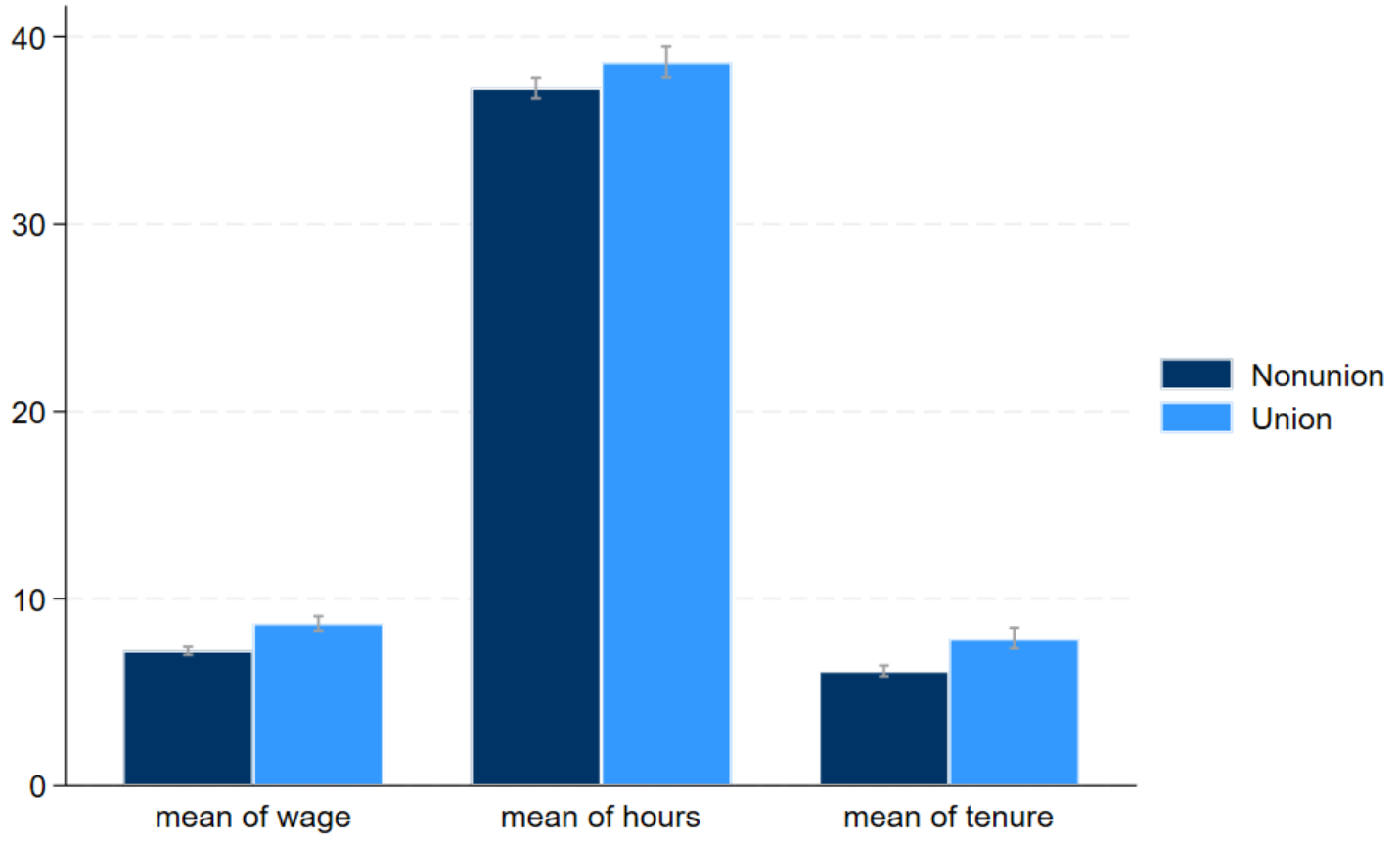
```
. frame create workers
. frame workers: webuse nlswork

. frame create companies
. frame companies: webuse grunfeld

. frames save myframeset, frames(workers companies)

. frame create RandD
. frame RandD: webuse xtcoint
. frames modify using myframeset, add(RandD)

. frames modify using myframeset, drop(workers)
```

New in graphics

- New `twoway` plots
  - `twoway heatmap`
  - `twoway rpcap`
  - `twoway rpspike`
- New statistics and options for `graph` plots
  - `graph bar (meanci), groupyvars blabel(, prefix() suffix())`
  - `graph dot (meanci), groupyvars`
  - `graph box, groupyvars`
- Other new options
  - `mlabprefix()` and `mlabsuffix()`
  - `legend(lastlabel())`
  - Hex-coded *colorstyle*

New in graphics

```
. graph bar (meanci) wage hours tenure, over(union) groupyvars
  bar(1, fcolor(#003366) lcolor(white)) bar(2, fcolor(#3399FF) lcolor(white))
```



Tables: Easier tabulations, exporting, and more

- [svy:] tabulate , collect()
- anova, oneway + collect
- table , export() title() titlestyles() notes() notestyles()
- collect notes , fortags()
- collect unget
- collect query layout
- collect style header, fvlevels()
- collect get: , commands()

New features in Stata 19

https://www.stata.com/new-in-stata/

- Machine learning via H2O: Ensemble decision trees
- Conditional average treatment effects (CATE)
- High-dimensional fixed effects (HDFE)
- Bayesian variable selection for linear model
- Interval-censored multiple-event Cox model
- Meta-analysis for correlations
- Bayesian bootstrap
- Control-function linear and probit models
- Bayesian quantile regression
- Inference robust to weak instruments
- Mundlak specification test
- Correlated random-effects (CRE) model
- Panel-data vector autoregressive (VAR) model
- SVAR models via instrumental variables
- Instrumental-variables local-projection IRFs
- Latent class model-comparison statistics
- Bayesian asymmetric Laplace model
- Do-file Editor: Autocompletion, templates, and more
- Graphics: Bar graph CIs, heat maps, and more
- Tables: Easier tabulations, exporting, and more
- Multiple datasets: Modify a set of frames
- Stata in French
- More

StataNow Updates

- Local average treatment effects (LATE)
- Power analysis for logistic regression
- VCE additions for linear models
- Do-file Editor change history ribbon
- Improved variable name truncation in the Data Editor
- Faster feasible generalized least squares estimation

Upcoming webinars

Conditional average treatment-effects estimation using Stata	October 15
Machine learning using Stata via H2O	November 13

Access past webinar recordings that use **lcstats**, new graphics features, and **stmgintcox**
www.stata.com/training/webinar/past-webinar-recordings/

Contact Stata's Technical Support team for further questions:
tech-support@stata.com