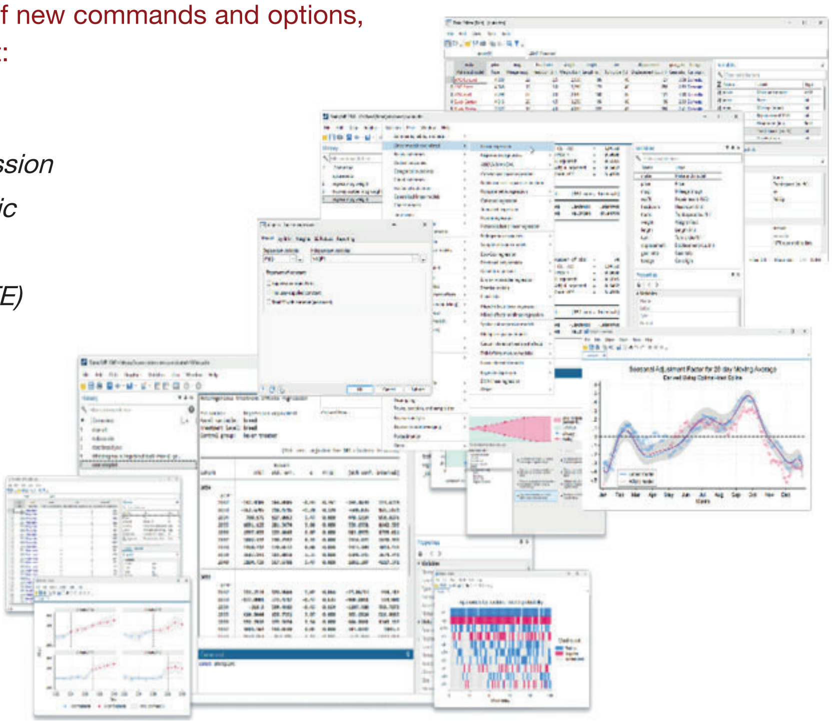
WHAT’S NEW!

Stata 19 contains the following exciting array of new commands and options, amongst which the following may be of interest:



- *Bayesian bootstrap and replicate weights*
- *Bayesian variable selection for linear regression*
- *Bayesian quantile regression via asymmetric Laplace likelihood*
- *Conditional average treatment effects (CATE)*
- *Control-function Linear and Probit models*
- *Correlated random-effects (CRE) model*
- *Do-file Editor:*
  - *Autocompletion*
  - *Templates*
  - *......*
- *Graphics:*
  - *Bar graph CIs*
  - *Heat Maps*
  - *......*
- *High-dimensional fixed effects (HDFE)*
- *Inference robust to weak instruments*
- *Instrumental-variables local-projection IRFs*
- *Latent class model-comparison statistics*
- *Machine learning via H2O: Ensemble decision trees*
- *Marginal Cox PH model for interval-censored multiple-events data*
- *Meta-analysis for correlations*
- *Mundlak specification test*
- *Panel-data vector autoregressive (VAR) model*
- *SVAR models via instrumental variables*
- *Tables:*
  - *Easier tabulations*
  - *Exporting*
  - *......*

Other new features include:

- *Alternative at-risk table for survival graphs*
- *Asymmetric Laplace likelihood for Bayesian models*
- *Bayesian predictions in user-defined evaluators*
- *Half-Cauchy and Rayleigh priors for Bayesian analysis*
- *Modify saved sets of frames*
- *PyStata enhancements*
- *Robust SEs for VAR models*
- *Stata in French*

www.tstat.it
www.tstattraining.eu
TStat - *StataCorp*'s Official Distributor serving:
Albania | Bosnia and Herzegovina | Croatia | Greece | Italy | Kosovo | North Macedonia | Malta | Montenegro | Serbia | Slovenja | Slovakia

BAYESIAN BOOTSTRAP AND REPLICATE WEIGHTS

Researchers can use the new **bayesboot** prefix to perform Bayesian bootstrap of statistics produced by official and community-contributed commands.

One can also use the new **rwgen** command and new options for the bootstrap prefix to implement specialized bootstrap schemes. **rwgen** generates standard replication and Bayesian bootstrap weights. **bootstrap** has new **fweights()** and **iweights()** options for performing bootstrap replications using the custom weights. **fweights()** allows users to specify frequency weight variables for resampling, and **iweights()** lets users provide importance weight variables.

These options extend **bootstrap**'s flexibility by allowing user-supplied weights instead of internal resampling, making it easier to implement specialized bootstrap schemes and enhance reproducibility. **bayesboot** is a wrapper for **rwgen** and **bootstrap** that generates importance weights using Dirichlet distribution and applies these weights when bootstrapping.

Bayesian bootstrap can be used to obtain more precise parameter estimates in small samples and incorporate prior information when sampling observations.

BAYESIAN VARIABLE SELECTION FOR LINEAR REGRESSION

The new **bayesselect** command provides a flexible Bayesian approach to identify the subset of predictors that are most relevant to users' outcome. It accounts for model uncertainty when estimating model parameters and performs Bayesian inference for regression coefficient.

A frequent problem in regression is identifying the subset of predictors that are most relevant to the outcome when users have many potential predictors. Variable selection, also called sparse regression, helps researchers with model interpretability and provides more stable inference.

Stata's Bayesian suite now includes a new command, **bayesselect**, that implements Bayesian variable selection for the linear model. **bayesselect** complements existing *Stata* commands related to variable selection, such as **lasso** and **bmaregress**.

bayesselect provides a flexible Bayesian approach to variable selection by using a variety of specially designed priors for coefficients, such as global–local shrinkage and spike-and-slab priors. **bayesselect** is fully integrated in *Stata*'s Bayesian suite and works seamlessly with all Bayesian postestimation routines.

As with other Bayesian regression procedures in *Stata*, posterior means, posterior standard deviations, Monte Carlo standard errors, and credible intervals of each predictor are reported for easy interpretation. Additionally, either inclusion coefficients or inclusion probabilities, depending on the selected prior, are included to indicate the importance of each predictor to model the outcome.

This variable-selection approach offers intuitive interpretation and stable inference.

BAYESIAN QUANTILE REGRESSION VIA ASYMMETRIC LAPLACE LIKELIHOOD

The new bayes: **qreg** command for quantile regression is now compatible with the **bayes** prefix. In the Bayesian framework, *Stata 19* combines the asymmetric Laplace likelihood function with priors to provide full posterior distributions for quantile regression coefficients.

In classical quantile regression, standard errors are computed by using bootstrap or kernel-based methods. In the Bayesian framework, posterior standard deviations play the role of standard errors.

By assuming a parametric likelihood model, the posterior standard deviations are estimated based on that model and may be more efficient.

Researchers can also use the asymmetric Laplace likelihood in **bayesmh** for random-effects quantile regression, simultaneous quantile regression, or to model nonnormal outcomes with pronounced skewness and kurtosis.

All implementations support standard Bayesian features, such as MCMC diagnostics, hypothesis testing, prediction.

CONDITIONAL AVERAGE TREATMENT EFFECTS (CATE)

Treatment effects estimate the causal effect of a treatment on an outcome. This effect may be constant or it may vary across different subpopulations. Researchers are often interested in whether and how treatment effects differ.

With the new **cate** command, users can go beyond estimating an overall treatment effect to estimating individualized or group-specific ones that address these types of research questions.

The **cate** command can estimate three types of CATEs: individualized average treatment effects, group average treatment effects, and sorted group average treatment effects. Beyond estimation, the **cate** suite provides features to predict, visualize, and make inferences about the CATEs.

The **cate** command is powerful, flexible and robust. It offers modeling of outcome and treatment models by offering lasso, generalized random forest (sometimes called honest forest), and parametric models. It provides two robust estimators (partialing out and augmented inverse probability weighting) to guard against machine learning mistakes, and it uses cross-fitting to avoid overfitting.

CONTROL-FUNCTION LINEAR AND PROBIT MODELS

The new **cfregress** and **cfprobit** commands allow users to fit control-function linear and probit models, which provide a flexible alternative to traditional instrumental variables (IV) methods for models with endogenous variables.

Users can include continuous, binary, fractional, and count endogenous variables and can easily test for endogeneity. Control-function models allow researchers to estimate causal relationships even when some explanatory variables are endogenous. Here first-stage models are fit for all endogenous variables and the residuals are then used to form control functions that are included in the main outcome model to account for endogeneity.

Researchers often use control-function methods when traditional IV methods cannot accommodate desired model features such as flexible handling of interacted endogenous variables or modeling endogenous binary, fractional, and count variables. The **cfregress** and **cfprobit** commands fit control-function models, allow for great flexibility in the interaction and modeling of endogenous variables, and provide standard errors that account for the inclusion of estimated control functions.

After fitting the model, users can easily perform tests of endogeneity.

First-stage models can be linear, Probit, Fractional probit, or Poisson, and their control functions can be interacted with other variables or with each other. Robust, cluster–robust, heteroskedasticity- and autocorrelation-consistent VCEs are allowed.

CORRELATED RANDOM-EFFECTS (CRE) MODEL

Easily fit CRE models to panel data with the new **cre** option of the **xtreg** command. Estimate coefficients for time-invariant regressors while getting the same coefficients for time varying regressors as those of **xtreg**, **fe**.

DO-FILE EDITOR

The Do-file Editor includes the following new features which make *Stata* coding more efficient and significantly faciltate the code writing task in *Stata*.

- Autocompletion of variable names, macros and stored results.
- Do-file Editor templates
- Do-file Editor current word and selection highlighting
- Bracket highlighting
- Code folding enhancements
- Do-file Editor temporary and permanent bookmarks
- Show whitespace and tabs
- Navigator panel
- and more

GRAPHICS

Stata 19 includes the following new graphics features, which have been repeatedly requested by *Stata* users. In particular:

- Heat maps
- Range and point plot with capped spikes
- Range and point plot with spikes
- Bar graphs with CIs, improved labeling, and control of bar groupings
- groupyvars
- Dot charts with CIs, improved labeling, and control of dot groupings.
- Box plots with improved labeling and control of box groupings.
- Colors by variable for more graphs
- and more

HIGH-DIMENSIONAL FIXED EFFECTS (HDFE)

Users can now absorb not just one, but multiple high-dimensional categorical variables in their linear regression, with or without fixed effects, and in linear models accounting for endogeneity using two-stage least squares. This is useful when users want their model to be adjusted for these variables, but estimating their effect is not of interest and is computationally expensive.

The **areg**, **xtreg**, **fe**, and **ivregress 2sls** commands now allow the **absorb()** option to be specified with multiple categorical variables. Previously, **areg** allowed only one variable in **absorb()**, while **xtreg**, **fe** and **ivregress 2sls** did not allow the option.

Absorbing high-dimensional categorical variables, rather than including indicators for them in users model, results in remarkable speed gains.

INFERENCE ROBUST TO WEAK INSTRUMENTS

Use the new **estat weakrobust** command to perform reliable inference on endogenous regressors. IV methods allow researchers to estimate causal relationships even when some explanatory variables are endogenous. IV methods exploit other variables—instrumental variables—that are correlated with the endogenous variables, but do not themselves suffer from endogeneity.

A well-known problem with IV methods in practice is that when instruments are only weakly correlated with the endogenous regressors, inference can become unreliable even in relatively large samples.

The new **estat weakrobust** postestimation command after **ivregress** performs Anderson–Rubin or conditional likelihood-ratio (CLR) tests on the endogenous regressors. **estat weakrobust** can also construct the associated confidence interval when there is only a single endogenous regressor. These tests and confidence intervals are fully robust to weak instruments.

This postestimation command supports all **ivregresss** estimators: **2sls**, **liml**, and **gmm**.

INSTRUMENTAL-VARIABLES LOCAL-PROJECTION IRFS

With the new **ivlpirf** command, users can account for endogeneity when using local projections to estimate dynamic causal effects.

Local projections are a method for estimating dynamic causal effects, which measure the effect of a shock to one variable on one or more outcomes over time. These causal effects are also called structural IRFs.

Local projections are used to estimate the effect of shocks on outcome variables. When the shock of interest is on an impulse variable that may be endogenous, **ivlpirf** can be used to estimate the IRFs, and the impulse variable may be instrumented using one or more exogenous instruments.

LATENT CLASS MODEL-COMPARISON STATISTICS

When one performs latent class analysis or finite mixture modeling, it is fundamental to determine the number of latent classes that best fits their data. With the new **lcstats** command, users can use statistics such as entropy and a variety of information criteria, as well as the Lo–Mendell–Rubin (LMR) adjusted likelihood-ratio test and Vuong–Lo–Mendell–Rubin (VLMR) likelihood-ratio test, to help users determine the appropriate number of classes.

The **lcstats** command offers options for specifying which statistics and test to report and to customize the look of the table.

MACHINE LEARNING VIA H2O: ENSEMBLE DECISION TREES

Machine learning methods are often used to solve research and business problems focused on prediction when the problems require more advanced modeling than linear or generalized linear models. Ensemble decision tree methods, which combine multiple trees for better predictions, are popular for such tasks. H2O is a scalable machine learning platform that supports data analysis and machine learning, including ensemble decision tree methods such as random forest and gradient boosting machine (GBM).

The new **h2oml** suite of *Stata* commands is a wrapper for H2O that provides end-to-end support for H2O machine learning analysis using ensemble decision tree methods. After using the **h2o** commands to initiate or connect to an existing H2O cluster, researchers can use the **h2oml** commands to perform GBM and random forest for regression and classification problems. The **h2oml** suite offers tools for hyperparameter tuning, validation, cross-validation, evaluating model performance, obtaining predictions, and explaining these predictions.

With the new **h2oml** suite, researchers can use machine learning via H2O to uncover insights from data when traditional statistical models fall short. Tune hyperparameters, use validation or cross-validation (CV), evaluate model performance, explain predictions, and more.

Stata users have long relied on linear regression, logistic regression, and traditional statistical models to uncover insights from their data. There are many applications, where relationships are often complex and nonlinear, and these classical methods may fall short in capturing more intricate data patterns.

- What if users predictors interact in ways linear (or logistic) regression models cannot capture?
- What if users model's accuracy plateaus despite careful variable selection?
- What if users need models that are robust against missing data and multicollinearity while also generalizing well beyond the scope of their observed data?
- And best of all, what if users can achieve all the above without the need to sacrifice explainability of predictions for predictive power?

This is where GBMs and random forest revolutionize the way users analyze their data in *Stata*. With seamless access to H2O's machine learning algorithms from within *Stata*, users can now harness the power of highperformance predictive models without leaving their familiar *Stata* environment. Users can simply use commands with intuitive *Stata* syntax to train sophisticated ensemble learning models that outperform traditional techniques.

MARGINAL COX PH MODEL FOR INTERVAL-CENSORED MULTIPLE-EVENTS DATA

Interval-censored multiple-event data commonly arise in longitudinal studies because each study subject may experience several types of events and those events are not observed directly, but are known to occur within some time interval. This data type arises in many fields, including medicine, epidemiology, biology, and sociology. For example, an epidemiologist studying chronic diseases might collect data on patients with multiple conditions, such as heart disease and metabolic disease, during different doctor visits. Similarly, a sociologist might conduct surveys to record major life events, such as job changes and marriages, at regular intervals. In ecology, researchers might monitor reproductive cycles of animals, such as nesting and birthing, through periodic observations. In these studies, researchers are often interested in evaluating the effects of certain factors on the event times. However, analyzing interval-censored multiple-event data is challenging because none of the event times are exactly observed and the dependence structure between different event times is often unknown.

Marginal proportional hazards models can be used to analyze interval-censored multiple-event data. These models do not require modeling the dependence structure between different events, thus providing more robust inference. They also produce parameters that can be interpreted as population-average effects. Additionally, they are faster than their random-effects counterparts.

The new **stmgintcox** command fits a marginal proportional hazards model to interval-censored multiple-event data. Researchers can use this command with either single or multiple-record-per-event data, and it supports TVCs for all events or specific ones.

After fitting the model, users can estimate and test the average effect of a covariate across all event times using a more powerful test than the classic multivariate Wald test. Users can also generate event-specific predictions, create plots of covariate-adjusted survivor and other functions, and produce goodness-of-fit plots after **stmgintcox**.

Users can now fit a marginal proportional hazards model for such data. The new **stmgintcox** command can accommodate single- and multiple-record-per-event data and supports time varying covariates for all events or specific ones.

META-ANALYSIS FOR CORRELATIONS

The **meta** suite now supports meta-analysis of correlation coefficients, allowing investigation of the strength and direction of relationships between variables across multiple studies. For instance, users may have studies reporting the correlation between education and income levels or between physical activity and improvements in mental health and wish to perform meta analysis.

Traditionally, meta analysis (MA) focuses on two-sample binary or continuous data, where the outcome of interest is measured across two groups often labeled as the treatment and control groups. For example, an MA may compare the efficacy of a new drug versus a placebo or the impact of two different educational interventions on student performance.

Users may however want to investigate the strength and direction of relationships between variables across multiple studies. This is where the MA of correlations comes into play. For example, a researcher might be interested in synthesizing findings from various studies to understand the correlation between digital device usage and sleep quality. Or an economist might conduct an MA to analyze the relationship between market volatility and investor behavior across countries from different studies. MA can be used in these cases to synthesize correlation coefficients from different studies.

Correlation studies are a cornerstone in many fields of research. Adding this feature makes **meta esize** one of the most flexible tools for meta-analysis available.

MUNDLAK SPECIFICATION TEST

Use the new **estat mundlak** postestimation command after **xtreg** to choose between random-effects (RE), fixedeffects (FE), or correlated random-effects (CRE) models even with cluster–robust, bootstrap, or jackknife standard errors.

Unlike the Hausman test for FE versus RE, the Mundlak test provides valid inference with cluster–robust, bootstrap, and jackknife standard errors.

PANEL-DATA VECTOR AUTOREGRESSIVE (VAR) MODEL

The new **xtvar** command allows users to fit a panel data vector autoregressive (VAR) model to analyze the trajectories of related variables when observe multiple units or panels over time.

VAR models have long been a staple of multivariate time-series analysis, but these models require relatively long series. Now users can apply the same tools to panel data, using observations across panels to compensate for the shorter span typical of these data. Users can evaluate the model using both moment- and model-selection criteria and Granger causality tests users are also able to interpret results using impulse–response functions (IRFs).

The new **xtvar** command has similar syntax and postestimation procedures as **var**, but it is appropriate for panel data rather than time-series data.

SVAR MODELS VIA INSTRUMENTAL VARIABLES

With the new **ivsvar** command, users can use instruments instead of short-run constraints to estimate dynamic causal effects.

Vector autoregressive (VAR) models describe how a collection of time-series variables interacts. In a VAR model, all variables are endogenous. When users want to estimate dynamic causal effects, they can place theoretical restrictions on the VAR model; these restrictions lead to structural VAR (SVAR) models. Here users focus on short-run SVAR models. Traditionally, these models place restrictions on how shocks impact the endogenous variables. Alternatively, if users have instruments, they can place restrictions on the relationships between shocks and instruments; this allows them to fit instrumental-variables (proxy) SVAR models. In these models, the impact of instrumented shocks (target shocks) on endogenous variables can be freely estimated.

ivsvar estimates the parameters of SVAR models by using instrumental variables. These estimated parameters can be used to trace out dynamic causal effects known as structural impulse–response functions (IRFs) using the familiar **irf** suite of commands. These IRFs describe how a shock to the SVAR model affects the model variables over time.

```
. irf set ivsvar.irf
. irf create model1
. irf graph sirf, impulse(shock)
```

For multiple instruments, use the minimum distance estimator with **ivsvar mdist**, and specify how the instruments are related to the target shocks.

TABLES

Stata 19 also includes includes a series of useful additions which allow users to more easily create and customize tables.

- Titles, notes, and exporting for tables
- Easier ANOVA tables
- Better labels with collect get
- Determine layout of a collection
- Control factor variables in headers
- Remove results from a collection
- Table-specific notes
- Tabulations with measures of association and tests

TStat S.r.l
via Rettangolo, 12/14 | 67039 Sulmona (AQ)
Tel +39 0864 210101 | tstat@tstat.it
www.tstat.eu | www.tstattraining.eu

Official Stata Trainer for EU
StataCorp's Official Distributor serving:
Albania, Bosnia and Herzegovina, Croatia,
Greece, Italy, Kosovo, North Macedonia, Malta,
Montenegro, Serbia, Slovenja, Slovakia

STATA/BE STATA/SE STATA/MP