Introduction to Stata 19

Presented by: **Anis Samet**

American University of Sharjah

timberlake
statistics | econometrics | forecasting

STATA 19

Agenda

- Why Stata?
- New in Stata 19

Why Stata

- Fast
- Accurate
- Easy to use
- Stata is a complete, integrated software package that provides all your data science needs—data manipulation, visualization, statistics, and automated reporting.

Why Stata

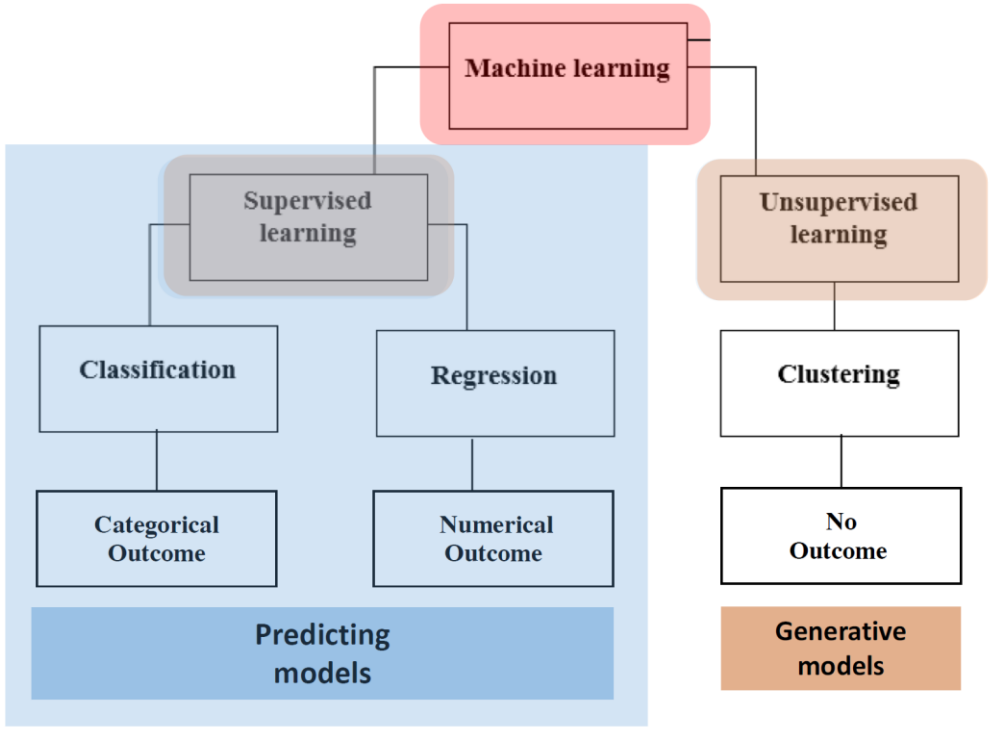
- Master your data
- Broad suite of statistical features
- Publication-quality graphics
- Automated reporting
- Truly reproducible research
- PyStata — Python integration
- Trusted and Reliable
- Continuously updated
- Easy to use, to automate, and to extend
- Advanced programming
- Community-contributed features
- Real documentation and World-class technical support
- Cross-platform compatible
- Widely used and Vibrant community

New in Stata 19

New features in Stata 19

- 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
- Bayesian quantile regression
- Panel-data vector autoregressive (VAR) model
- Correlated random-effects (CRE) model
- Bayesian bootstrap
- Control-function linear and probit models
- SVAR models via instrumental variables
- Instrumental-variables local-projection IRFs
- Latent class model-comparison statistics
- Bayesian asymmetric Laplace model
- Inference robust to weak instruments
- Meta-analysis for correlations
- Mundlak specification test
- 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

Introduction to Machine Learning (ML)

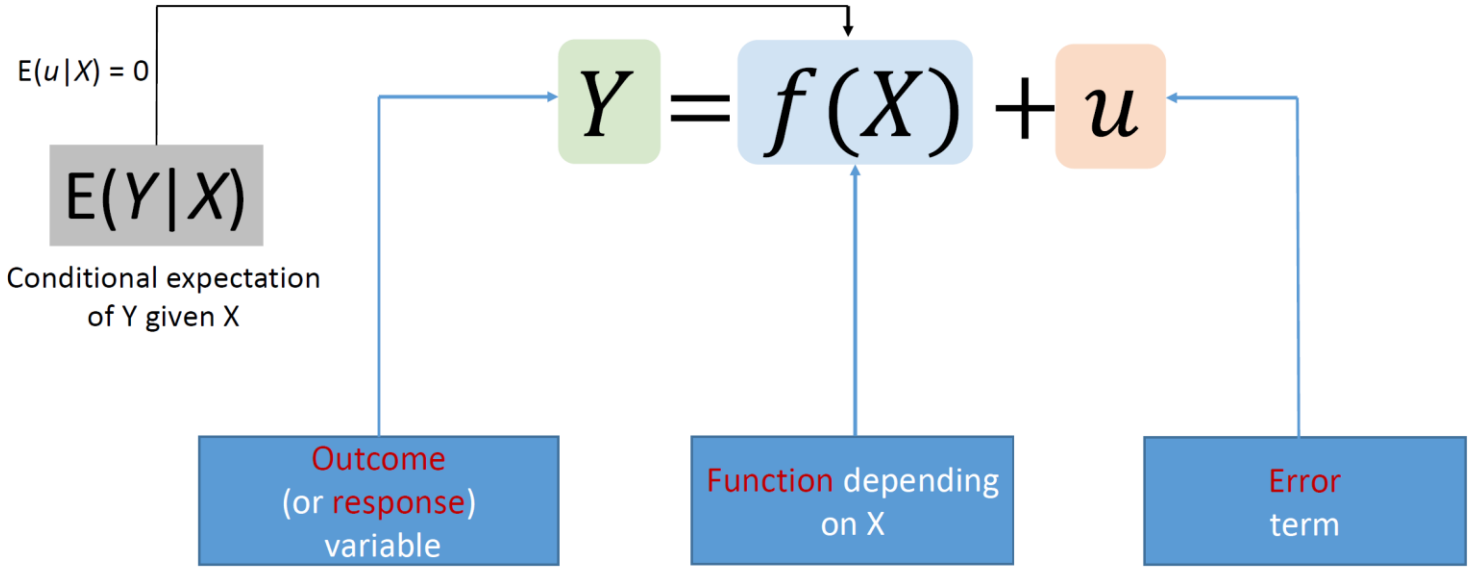


Source: Cerulli (2023)

Introduction to Machine Learning (ML)

Modelling as "learning"

More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form



Source: Cerulli (2023)

ML Techniques in Stata

- Several Stata command can be used in ML:
- rforest
- svm
- boost (only for Windows)
- sctree, srtree
- subset
- mlp2
- Python related commands: r_ml_stata_cv and c_ml_stata_cv

Machine learning via H2O: Ensemble decision trees

- With the new h2oml suite in Stata 19, use machine learning via H2O to uncover insights from data when traditional statistical models fall short
- Use ensemble decision trees—gradient boosting machine (GBM) and random forest—to perform classification or regression.
- With the access to H2O's machine learning algorithms from within Stata, you can now harness the power of high-performance predictive models without leaving your familiar Stata environment
- With tools such as Shapley additive explanations (SHAP) values, partial dependence plots (PDPs), and variable importance rankings, GBM and random forest provide powerful predictions while maintaining explainability—no tradeoffs needed

Machine learning via H2O: Ensemble decision trees

- The GBM algorithm works in a sequential manner, where each model in the sequence is trained to correct the mistakes of the previous model. This is achieved by focusing on the pseudo-residuals (the difference between the actual and predicted values) from the previous model
- Random Forest is a type of machine learning algorithm that helps us make predictions based on lots of decision trees. A decision tree is like a flowchart where you ask questions to split data into smaller groups, helping you predict an outcome (like yes or no, or a number)

Machine learning via H2O: Ensemble decision trees

[H2OML] h2oml — Introduction to commands for Stata integration with H2O machine learning
(View complete PDF manual entry)

Description

This entry describes commands for performing predictive analysis using H2O machine learning methods, specifically ensemble decision tree methods, in Stata. H2O is a scalable and distributed machine learning and predictive analytics platform that allows you to perform data analysis and machine learning. It provides parallelized implementations of many widely used supervised and unsupervised machine learning methods. For more details, see **[H2OML] H2O setup**, **[P] H2O intro**, and https://www.stata.com/h2o/h2o18/h2o_intro.html#what-is-h2o. For a software-free introduction to machine learning, see **[H2OML] Intro**.

Supervised learning

h2oml gbm	gradient boosting machine
h2oml gbregress	gradient boosting regression
h2oml gbbinclass	gradient boosting binary classification
h2oml gbmulticlass	gradient boosting multiclass classification
h2oml rf	random forest
h2oml rfregress	random forest regression
h2oml rfbinclass	random forest binary classification
h2oml rfmulticlass	random forest multiclass classification

Estimation results and postestimation frame

h2omlest	catalog H2O estimation results
h2omlpostestframe	specify frame for postestimation analysis

Conditional average treatment effects (CATEs)

- The new **cate** command goes beyond estimating overall treatment effects in your analysis of causal effects to estimating individualized and group-specific ones. Compare different interventions and policies. Explore treatment-effects heterogeneity
- **cate** helps us answer questions such as the following:

1. Are the treatment effects heterogeneous?
2. How do the treatment effects vary with some variables?
3. Do the treatment effects vary across prespecified groups?
4. Are there unknown groups in the data for which treatment effects differ?
5. Which is best among possible treatment-assignment rules?

Conditional average treatment effects (CATEs)

- **cate** estimates conditional average treatment effects (CATEs), which are average treatment effects (ATEs) conditional on a set of variables for which the treatment effects may vary. Estimating CATEs allows us to study treatment-effect heterogeneity and evaluate treatment-assignment policies
- **cate** provides three different CATE estimates: individualized average treatment effects (IATEs), group average treatment effects (GATEs), and sorted group average treatment effects (GATESs)

[CAUSAL] cate — Conditional average treatment-effects estimation
(View complete PDF manual entry)

Syntax

Partialing-out estimator

cate po (*ovar catevarlist*) (*tvar*) [*if*] [*in*] [, *options*]

Augmented inverse-probability weighting estimator

cate aipw (*ovar catevarlist*) (*tvar*) [*if*] [*in*] [, *options*]

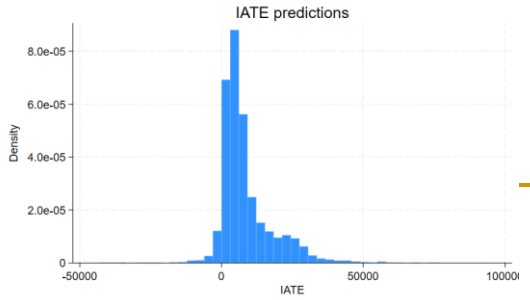
ovar is a continuous outcome of interest.

catevarlist specifies the covariates of the CATE model -- the conditioning variables for the treatment effects. *catevarlist* may contain factor variables; see fvvarlist.

tvar must be a binary variable representing the treatment levels.

Conditional average treatment effects (CATEs)

- **Stata example:**
- ***We open the assets3 dataset. We define a global macro, catecovars, that stores the names of the variables on which we will condition
- webuse assets3
- global catecovars age educ i.(incomecat pension married twoearn ira ownhome)
- **estimate the effect of 401(k) eligibility (e401k) on net financial assets (asset) to answer the following question
- cate po (assets $catecovars) (e401k)
- ***use categraph histogram to draw a histogram of the predicted IATEs and see their distribution
- categraph histogram
- **To test whether the treatment effects are heterogeneous, use estat heterogeneity
- estat heterogeneity
- ***We use categraph iateplot to visually examine the change in the IATE function across educ levels
- categraph iateplot educ



High-dimensional fixed effects (HDFE)

- Linear models with high-dimensional categorical variables
- Absorb multiple high-dimensional categorical
- Absorb not just one but multiple high-dimensional categorical variables in your linear, fixed-effects linear, and instrumental-variables linear models using option absorb() with commands areg, xtreg, fe, and ivregress 2sls
- Similar to **reghdfe** Stata user-written command

Title

reghdfe — Linear regression with multiple fixed effects. Also supports individual FEs with group-level outcomes

Syntax

Least-square regressions (no fixed effects):

reghdfe *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, options]

Fixed effects regressions:

reghdfe *depvar* [*indepvars*] [*if*] [*in*] [*weight*] , **absorb**(*absvars*) [options]

Fixed effects regressions with group-level outcomes and individual FEs:

reghdfe *depvar* [*indepvars*] [*if*] [*in*] [*weight*] , **absorb**(*absvars indvar*) **group**(*groupvar*) **individual**(*indvar*) [options]

High-dimensional fixed effects (HDFE)

- **Stata example:**
- *Import panle data
- webuse nlswork
- xtset id year
- ****Run linear regression without clustering
- reg ln_w tenure age
- ****Run linear regression with double clustering
- areg ln_w tenure age, absorb (race msp)
- **Run panel regression with double clustering
- xtreg ln_w tenure age, fe absorb (race msp)

Correlated random-effects (CRE) model

- Want coefficient estimates of time-invariant covariates in your panel-data model? Fit a random-effects model. Want to allow for correlation between covariates and unobserved panel-level effects
- We study the effect on wages of time-varying variables, such as age or tenure. At the same time, we are interested in the effects of time-invariant variables, such as race
- An FE model will omit any variable that remains constant across time and thus cannot fully answer our research question
- An RE model may yield inconsistent estimates because of the possible correlation between individual time-invariant heterogeneity and the regressors age and tenure

Correlated random-effects (CRE) model

- *Import panel data
- webuse nlswork
- *********Run CRE regression
- xtreg ln_wage tenure age i.race, cre vce(cluster idcode)

Marginal Cox proportional hazards model for interval-censored multiple-event data

- Need to analyze event times from multiple types of events such as the onset of diabetes and hypertension? Don't know the exact event times?
- The new **stmgintcox** command analyzes such interval-censored multiple-event data and account for possible correlation between event times across the different events
- Stata 18 expanded the functionality of stintcox to support time-varying covariates (TVCs)
- The new **stmgintcox** command fits a marginal proportional hazards model to interval-censored multiple-event data. You can use this command with either single- or multiple-record-per-event data, and it supports TVCs for all events or specific ones
- It also offers flexible ways to specify models with event-specific covariates

Marginal Cox proportional hazards model for interval-censored multiple-event data

- Interval-censored multiple-event data (or, more precisely, interval-censored event-times data on multiple events) 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
- 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.
- 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 Cox proportional hazards model for interval-censored multiple-event data

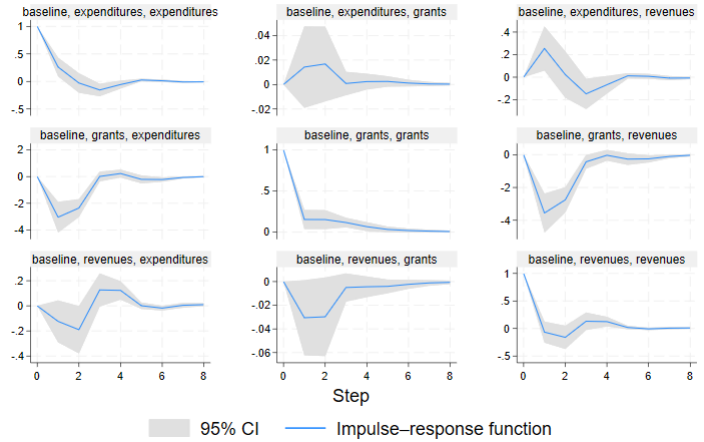
- **Stata example:**
- ****Import data
- webuse aric
- list id event ltime rtime bmi - diabp if id==91 | id==92, sepby(id) noobs
- *******Run marginal proportional hazards model to interval-censored multiple-event data
- stmgintcox age i.male i.community i.race bmi glucose sysbp diabp, id(id) event(event) interval(ltime rtime)

Panel-data vector autoregressive model

- With the new **xtvar** command, you can now fit a panel-data vector autoregressive (VAR) model to analyze the trajectories of related variables when you 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
- We can apply the same tools to panel data, using observations across panels to compensate for the shorter span typical of these data
- We can evaluate the model using moment- and model-selection criteria and Granger causality tests
- We can interpret results using impulse–response functions (IRFs)

Panel-data vector autoregressive model

- **Stata example:**
- ****Import data
- webuse swedishgov.dta
- ***We specify the lags(2) option with the xtvar command to include two lags of each dependent variable in each equation
- xtvar grants revenues expenditures, lags(2)
- *** Perform a Granger causality test to see whether grants Granger causes expenditures
- vargranger
- ***obtain IRFs after fitting a panel-data VAR model
- irf create baseline, set(swedish_govt_irfs)
- irf graph irf, byopt(yrescale)



Inference robust to weak instruments

- Do you have weak instruments in your instrumental-variables (IV) regression?
- 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

Inference robust to weak instruments

- **Stata example:**
- ****Import data
- webuse hsng
- ***Instrumental variables 2SLS regression
- ivregress 2sls rent pcturban (hsngval = i.region), vce(robust)
- ***If we suspect that our instruments for hsngval are weak, we can perform a test on the coefficient of hsngval that is robust to weak instruments
- estat weakrobust

- We can also use the user-written command ivreg2

Meta-analysis for correlations

- Traditionally, 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
- Sometimes, we may want to investigate the strength and direction of relationships between variables across multiple studies
- This is where the MA of correlations comes into play

Meta-analysis for correlations

- **Stata example:**
- webuse pupiliq
- meta set stdmdiff se
- meta summarize
- meta forestplot

Do-file Editor enhancements

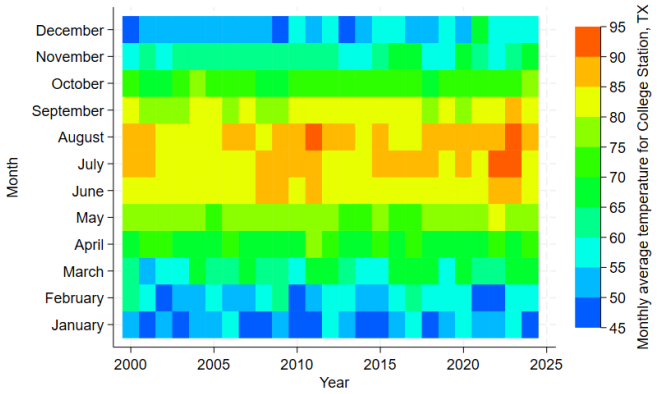
- The Do-file Editor now provides an even better and more customizable environment for coding in Stata with additional autocompletion, templates, improved code folding, and more
- Autocompletion of variable names, macros, and stored results: if you pause briefly as you type, suggestions of variable names from data in memory, macros, and stored results will appear in addition to the command names and existing words that appeared previously
- Templates: By popular request, you can now save time and ensure consistency when you create new documents
- Current word and selection highlighting: The Do-file Editor will now highlight all case-insensitive occurrences of the current word under the cursor
- Show whitespace and tabs
- Navigator panel

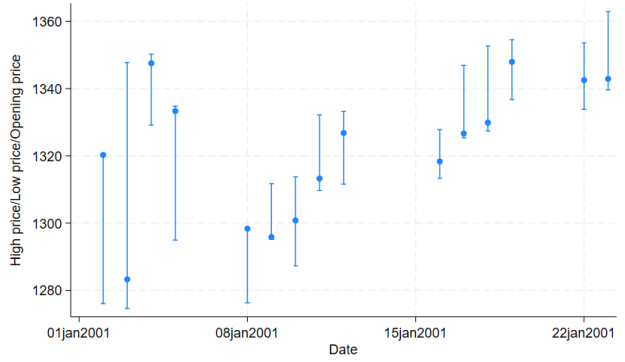
New graphics features

- Stata 19 supports a new two-way plottype heatmap to create a heat map
- Stata 19 also has two other new two-way plottypes, **rpcap** and **rpspike**, which plot a value and range, such as high, low, and opening daily stock prices. To connect the high and low values, rpcap uses a capped spike, and rpspike uses a spike. Both use a marker for the opening value.
- Additionally, new features have been added to graph bar, graph dot, and graph box

New graphics features

- **Stata example:**
- ****Import data
- webuse cstemp
- ********Twoway heat map
- twoway heatmap temp month year, ylabel(1(1)12, valuelabel) ccuts(45(5)95)
- ****Import data
- sysuse sp500
- ****Plot high low open
- twoway rpcap high low open date in 1/15
- twoway rpspike high low open date in 1/15





New reporting features

- Building customizable and reproducible tables is now even easier in Stata 19
- The **putdocx and putexcel** table commands for creating and customizing tables in a single command now allows you to add a title and notes and export tables.
- It is also easier to create tables from a collection
- anova, oneway, and tabulate now work seamlessly with collect. And collect now allows you to query layout specifications and remove results from a collection
- Also refer to **tabstat2excel** and **outreg2** Stata user-written commands

New reporting features

- **Highlights**
- New in **putdocx**
  - Include bookmarks in paragraphs and tables
  - Include alternative text to be read by voice software for images
  - Include Scalable Vector Graphics (.svg) images
- New in **putexcel**
  - Freeze a worksheet at a specific row or column
  - Insert a page break at a specific row or column
  - Insert a header and footer into a worksheet
  - Include hyperlinks in cells
  - Create a named cell range
- https://www.stata.com/new-in-stata/new-reporting-features/

New reporting features

- **Stata example:**
- ******Export into word document
- ****Import data
- webuse nhanes2l, clear
- putdocx begin
- quietly: dtable age weight bpsystol tcresult tgresult i.sex, by(diabetes, tests) title(Table 1) continuous(age weight, test(none)) factor(sex, test(none)) nformat(%6.1f mean sd)
- collect style putdocx, layout(autofitcontents)
- putdocx collect
- putdocx save bwtreport, replace

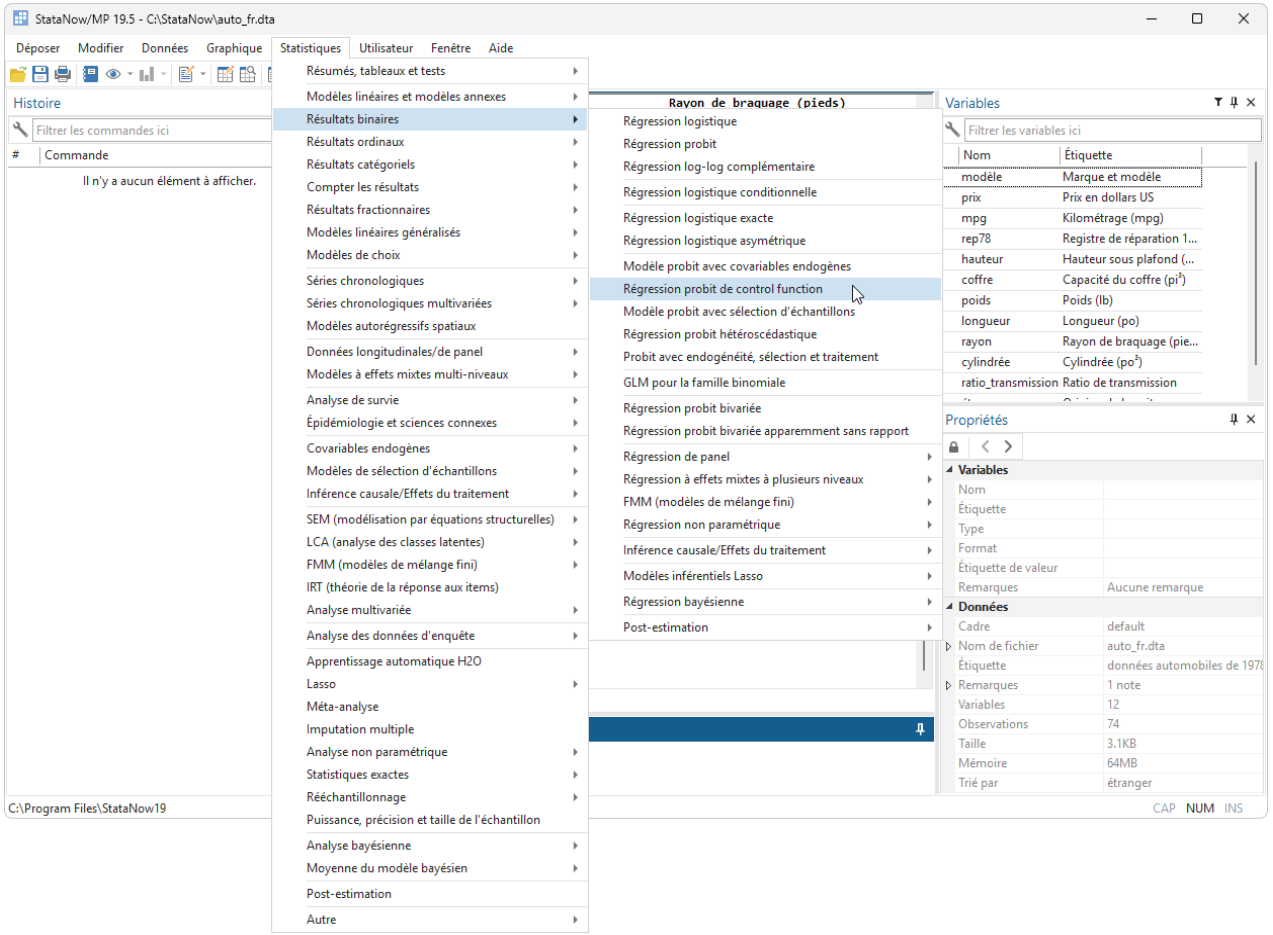
New reporting features

- **Stata example:**
- **********Export into Excel document
- ****Import data
- webuse census, clear
- foreach x of varlist pop death marriage {
- bysort region: egen avg_`x' = mean(`x')
- }
- export excel region avg_* pop death marriage state using report2.xlsx, firstrow(variables) replace
- putexcel set report2.xlsx, modify
- putexcel sheetset, split(1, 4)
- putexcel save

New reporting features

- **Stata example:**
- *********Export Tables into word and Excel document
- ****Import data
- clear
- webuse nhanes2l
- table hlthstat, statistic(frequency) statistic(percent) missing title(Table 1) note("Source: NHANES II") note("Sample includes blank but applicable and missing responses.") export(table.docx)
- table hlthstat, statistic(frequency) statistic(percent) missing title(Table 1) note("Source: NHANES II") note("Sample includes blank but applicable and missing responses.") export(table2.xlsx)

Stata in French



References

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

timberlake
data science | statistics | econometrics

THANKS !

We remain at your disposal to answer any questions .
your Stata-related requests .

E-MAIL

info@timberlake.ae

PHONE

+971 4 431 8123

STATA®