2022 Swiss Stata Conference

Room 205, Hallerstrasse 6, 3012 Bern
University of Bern

https://www.stata.com/meeting/switzerland22/

November 18, 2022

Program

08:30 Registration

08:50 Welcome

09:00 **pystacked: Stacking generalization and machine learning in Stata**

Achim Ahrens (ETH Zurich), Christian B. Hansen (University of Chicago), and Mark E. Schaffer (Heriot-Watt University)

`pystacked` implements stacked generalization (Wolpert, 1992) for regression and binary classification via Python's `scikit-learn`. Stacking combines multiple supervised machine learners—the "base" or "level-0" learners—into a single learner. The currently supported base learners include regularized regression, random forest, gradient boosted trees, support vector machines, and feed-forward neural nets (multi-layer perceptron). `pystacked` can also be used with as a 'regular' machine learning program to fit a single base learner and, thus, provides an easy-to-use API for `scikit-learn`'s machine learning algorithms.

09:25 **ddml: Double/debiased machine learning in Stata**

Achim Ahrens (ETH Zurich), Christian B. Hansen (University of Chicago), Mark E. Schaffer (Heriot-Watt University), and Thomas Wiemann (University of Chicago)

We introduce the Stata package `ddml` which implements Double/Debiased Machine Learning (DDML) for causal inference aided by supervised machine learning. Five different models are supported, allowing for multiple treatment variables in the presence of high-dimensional controls and/or instrumental variables. `ddml` is compatible with many existing supervised machine learning programs in Stata.

09:50 **Stata-Python API for bulk data download: Example with UN Comtrade**

Ka Lok Wong (Steve) (Geneva Graduate Institute)

This presentation aims to guide the audience through the bulk download of Comtrade data via a Stata-Python integration setup, which has been made available since Stata 16. Though this presentation is explicitly about the UN Comtrade dataset, the methodology employed is generalizable to other data platforms that allow API downloads.

The UN Comtrade Database is one of the best sources when it comes to bilateral trade data by product code. As of early 2022, it covers more country-year observations than the World Trade Organization and the International Trade Centre. However, tailoring the raw data to each researcher's needs is often time-consuming. Using the Comtrade API with my Stata-Python setup would allow researchers to tailor their downloaded data to their desired specification. In addition, employing this setup significantly reduces human error when compared to the manual downloading and cleaning of Comtrade data.

Full blog post can be found here: `https://www.stevekwong.com/blog/apicomtrade`

10:05 Break

10:35 Flexible and fast estimation of quantile treatment effects: The rqr and rqrplot commands

Nicolai T. Borgen (University of Oslo), Andreas Haupt (Karlsruhe Institute of Technology), and Øyvind Wiborg (University of Oslo)

Using quantile regression models to estimate quantile treatment effects is becoming increasingly popular. This paper introduces the `rqr` command that can be used to estimate residualized quantile regression (RQR) coefficients and the `rqrplot` postestimation command that can be used to effortless plot the coefficients. The main advantages of the `rqr` command compared to other Stata commands that estimate (unconditional) quantile treatment effects are that it can include high-dimensional fixed effects and that it is considerably faster than the other commands.

11:00 Stata commands to estimate quantile regression with panel and grouped data

Blaise Melly (Unversity of Bern) and Martina Pons (Unversity of Bern)

In this presentation, we introduce two Stata commands that allow estimating quantile regression with panel and grouped data. The commands implement two-step minimum-distance estimators. We first compute a quantile regression within each unit and then apply GMM to the fitted values from the first stage. The command `xtmdqr` applies to classical panel data, where we follow the same units over time while the command `mdqr` applies to grouped data, where the observations are at the individual level, but the treatment varies at the group level. Depending on the variables assumed to be exogenous, this approach provides quantile analogs of the classical least squares panel data estimators such as the fixed effects, random effects, between, and Hausman-Taylor estimators. For grouped (instrumental) quantile regression, we provide a more precise estimator than the existing estimators. In our companion paper (Melly and Pons, "Minimum Distance Estimation of Quantile Panel Data Models"), we study the theoretical properties of these estimators.

11:25 Improved Tests for Granger Non-Causality in Panel Data

Jiaqi Xiao (University of Birmingham), Arturas Juodis (University of Amsterdam), Yiannis Karavias (University of Birmingham), Vasilis Sarafidis (BI Norwegian Business School), and Jan Ditzen (Free University of Bozen-Bolzano)

Granger causality is an important aspect of applied panel (longitudinal) data analysis as it can be used to determine whether one variable is useful in forecasting another. This presentation describes `xtgranger`, a community-contributed Stata command, which implements the panel Granger non-causality test of Juodis, Karavias, and Sarafidis (2021). This test offers superior

size and power performance to existing tests, which stems from the use of a pooled estimator that has a faster convergence rate. The test has several other useful properties; it can be used in multivariate systems, it has power against both homogeneous as well as heterogeneous alternatives, and it allows for cross-section dependence and cross-section heteroskedasticity. The command is used to examine the type of temporal relation between profitability, cost efficiency and asset quality in the U.S. banking industry.

11:50 Drivers of COVID-19 deaths in the United States: A two-stage modeling approach

Kit Baum (Boston College), Andrés Garcia-Suaza (University del Rosario), Miguel Henry (Greylock McKinnon Associates), and Jesús Otero (University del Rosario)

We offer a two-stage (time-series and cross-section) econometric modeling approach to examine the drivers behind the spread of COVID-19 deaths across counties in the United States. Our empirical strategy exploits the availability of two years (January 2020 through January 2022) of daily data on the number of confirmed deaths and cases of COVID-19 in the 3,000 U.S. counties of the 48 contiguous states and the District of Columbia. In the first stage of the analysis, we use daily time-series data on COVID-19 cases and deaths to fit mixed models of deaths against lagged confirmed cases for each county. Because the resulting coefficients are county specific, they relax the homogeneity assumption that is implicit when the analysis is performed using geographically aggregated cross-section units. In the second stage of the analysis, we assume that these county estimates are a function of economic and sociodemographic factors that are taken as fixed over the course of the pandemic. Here we employ the novel one-covariate-at-a-time variable-selection algorithm proposed by Chudik et al. (Econometrica, 2018) to guide the choice of regressors.

12:15 Lunch

13:15 Bayesian Time Series in Stata 17

David Schenck (Senior Econometrician at StataCorp)

Stata 17 introduced Bayesian support for several multivariate time-series commands. In this talk, I will discuss Bayesian vector autoregressive models and Bayesian DSGE models. Bayesian estimation is well suited to these models because economic considerations often impose structure that is captured well by informative priors. I will describe the main features of these commands, as well as Bayesian diagnostics, posterior hypothesis tests, predictions, impulse-response functions, and forecasts.

14:15 Break

14:35 Network regressions in Stata

Jan Ditzen (Free University of Bozen-Bolzano), William Grieser (Texas Christian University), and Morad Zekhnini (Michigan State University)

Network analysis has become critical to the study of social sciences. While several Stata programs are available for analysing network structures, programs that execute regression analysis with a network structure are currently lacking. We fill this gap by introducing the `nwxtregress` command. Building on spatial econometric methods (LeSage and Pace 2009), `nwxtregress` uses MCMC estimation to produce estimates of endogenous peer effects, as well as own-node (direct) and cross-node (indirect) partial effects, where nodes correspond to cross-sectional units of observation, such as firms, and edges correspond to the relations between nodes. Unlike existing spatial regression commands (for example, `spxtregress`), `nwxtregress` is designed to handle unbalanced panels of economic and social networks as in Grieser et al. (2021). Networks can be directed or undirected with weighted or unweighted edges, and

they can be imported in a list format that does not require a shapefile or a Stata spatial weight matrix set by `spmatrix`. Finally, the command allows for the inclusion or exclusion of contextual effects. To improve speed, the command transforms the spatial weighting matrix into a sparse matrix. Future work will be targeted toward improving sparse matrix routines, as well as introducing a framework that allows for multiple networks.

15:00 Exchangeably weighted bootstrap schemes

Philippe Van Kerm (Luxembourg Institute of Socio-Economic Research and University of Luxembourg)

The exchangeably weighted bootstrap is one of the many variants of bootstrap resampling schemes. Rather than directly drawing observations with replacement from the data, weighted bootstrap schemes generate vectors of replication weights to form bootstrap replications. Various ways to generate the replication weights can be adopted and some choices bring practical computational advantages. This talk demonstrates how easily such schemes can be implemented and where they are particularly useful, and introduces the `exbsample` command which facilitates their implementation.

15:25 Marginal odds ratios: What they are, how to compute them, and why applied researchers might want to use them

Ben Jann (University of Bern) and Kristian Brent Karlson (University of Copenhagen)

Logistic response models form the backbone of much applied quantitative research in epidemiology and the social sciences. However, recent methodological research highlights difficulties in interpreting odds ratios, particularly in a multivariate modeling setting. These difficulties arise from the fact that coefficients from nonlinear probability models such as the logistic response model (i.e., log odds ratios) depend on model specification in ways that differ from the linear model. Applied researchers have responded to this situation by reporting marginal effects on the probability scale implied by the nonlinear probability model or obtained by the linear probability model.

Although marginal effects on the probability scale have many desirable properties, they do not align well with research in which relative inequality is a key concept. We argue that in many cases the odds ratio is preferable because it is a relative measure that does not depend on the marginal distribution of the dependent variable. In our talk, we aim to remedy the declining popularity of the odds ratio by introducing what we term the "marginal odds ratio"; that is, logit coefficients that have similar properties as marginal effects on the probability scale, but which retain the odds ratio interpretation. We define the marginal odds ratio theoretically in terms of potential outcomes, both for binary and continuous treatments, we develop estimation methods using three different approaches (G-computation, inverse probability weighting, RIF regression), and we present examples that illustrate the usefulness and interpretation of the marginal odds ratio.

15:50 It is all about the data

Maarten Buis (University of Konstanz)

This talk is a collection of tips for exploring a new dataset and preparing a dataset using both official and community contributed commands. Community contributed commands that will be covered are `lany`, `lookfor2`, `htmlcb`, and `closedesc`.

16:15 Break

16:45 btable: Extensive summary tables in Stata

Lukas Bütikofer (University of Bern)

The construction of summary tables is a very common, repetitive and time-consuming step in data analysis. `btable` is a flexible, easy-to-use and powerful algorithm for generating such tables in Stata. It is freely available from github.

`btable` can summarize continuous, categorical, count and time-to-event variables within one table using various descriptive statistics, which can be individually chosen and combined for each variable. If the summary is grouped, effect measures with confidence intervals and p-values are added. User-defined effect measures and tests can be integrated.

The table is constructed in a two-step approach using two functions: `btable` produces an unformatted, raw table, which is then formatted by `btable_format` to produce a final, publication-ready table. By default, the raw table contains all descriptive statistics, and, if grouped, effect measures with confidence intervals and p-values. The formatting step allows for a variable-specific selection and formatting.

The two-step approach separates data analysis and formatting. The analysis step does not change the current dataset and the raw data table can be loaded, formatted by hand, or used for other purposes. The formatting step can be modified without re-running the analysis.

17:00 Visualizing categorical data with hammock plots

Matthias Schonlau (University of Waterloo)

Visualizing data with more than two variables is not straight forward, especially when some variables are categorical rather than continuous. My hammock plots are one option to visualize categorical data and mixed categorical/continuous data. Hammock plots can be viewed as a generalization of parallel coordinate plots where the lines are replaced by rectangles that are proportional to the number of observations they represent. I will introduce my Stata program for hammock plots and give several short examples where I have found them useful.

17:25 circlebar: A Stata package for plotting circular bar graphs

Asjad Naqvi (Austrian Institute for Economic Research and Vienna University of Economics and Business)

The presentation will introduce `circlebar`, a Stata package that allows users to visualize data as circular bar graphs organized in polar coordinates. The command allows for flexibility of selecting and changing bar dimensions including starting and ending circles, colors and label placements, and controlling spacing between the bars.

17:50 Open panel discussion with Stata developers

18:20 End of conference

19:00 Conference dinner (Lötschberg, Zeughausgasse 16)