XVII ITALIAN STATA USERS CONFERENCE

Florence, 19-20 May 2022

PROGRAM

8.45 - 9.00 Registration

9.00 - 10.00 SESSION I - EXPLOITING THE POTENTIAL OF STATA 17, I

Custom estimation tables • Jeff Pitblado, Executive Director of Statistical Software, StataCorp

This presentation illustrates how to construct custom tables from one or more estimation command. I demonstrate how to add custom labels for significant coefficients and make targeted style edits to cells in the table using the following commands:

- collect get
- collect dir
- collect dims
- collect levelsof
- collect label list
- collect label values
- collect layout
- collect query header
- collect style header
- collect style showbase
- collect style row
- collect style cell
- collect query column
- collect style column
- collect style stars
- collect query column
- collect preview
- etable

I begin with a description of what constitutes a collection, and how items (numeric and string results) in a collection are tagged (identified) and conclude with a simple workflow to enable users to build their own custom tables from estimation commands. This talk motivates the construction of estimation tables and concludes with the convenience command **-etable-**.

10.00 - 11.15 SESSION II - COMMUNITY CONTRIBUTED, I

Machine Learning using Stata/Python • Giovanni Cerulli, National Research Council - IRcRES, Rome

Two related Stata modules, **r_ml_stata** and **c_ml_stata**, are presented for fitting popular machine learning (ML) methods both in regression and classification settings. Using the recent Stata/Python integration platform( ) introduced in Stata 16, these commands provide hyper-parameters' optimal tuning via K-fold cross-validation using grid search. More specifically, they make use of the Python Scikit-learn API to carry out both cross-validation and outcome/label prediction.

SCIENTIFIC COMMITTEE

Una-Louise BELL
Rino BELLOCCO
Giovanni CAPELLI
Maurizio PISATI

CONTACT

Monica GIANNI
Via Rettangolo, 12-14
67039 Sulmona (AQ)
T. +39 0864 210101

www.tstat.it | www.tstattraining.eu
formazione@tstat.it | training@tstat.eu

A Stata routine for estimating the blocking with regression adjustment • Martina Bazzoli, FBK-IRVAPP, Research Institute for the Evaluation of Public Policies, Trento

psreg command implements the blocking with regression adjustments estimator, proposed by Imbens (J. Human Resources, 2015). It relies on the estimate of the propensity score and uses regressions in subclasses (blocks) of the propensity score. The ATT is given by estimates within-block averaged for the number of treated units in each block. In the case of ATE the estimates are averaged for the number of units (treated and untreated) in each block.

11.15 - 11.30 Coffee break

11.30 - 13.00 SESSION III - COMMUNITY CONTRIBUTED, II

A Stata package for Cluster Weighted Modeling • Daniele Spinelli, University of Milan–Bicocca

The *Cluster-Weighted Model (CWM)* is a member of the family of the Mixtures of Regression Models and it is also referred in literature to as Mixture of Regression with Random Covariates. These models extend finite mixture models by allowing the researcher to model the marginal distribution of regression covariates along with the conditional distribution. The attention on CWMs is increasing; indeed, software for estimating this kind of models is available to R users but not for Stata users. Thus, the aim of this paper is to introduce the Stata package **cwmglm**. This package extends the capabilities of **fmm** by introducing more advanced mixture models based on maximum likelihood estimation and the expectation maximization EM algorithm.

cwmglm allows users to fit CWMs based on the most common generalized linear models (GLM) with random covariates. The supported GLM families are Gaussian, Poisson and binomial, while the allowed marginal distributions for the covariates are multivariate normal, multinomial, binomial, and Poisson. **Cwmglm** extends the current capabilities in the estimation of CWMs by allowing users to evaluate model fit by introducing the generalized determination coefficients and by incorporating bootstrap-based inference. These features are not available in the current version of the R-package software for CWMs. Furthermore, **cwmglm** allows one to estimate parsimonious models of Gaussian distributions. This approach is based on assuming the correlation structure between concomitants within multivariate Gaussian mixture components and on the equality/inequality of variance-covariance matrices between components. Fourteen parsimonious models are possible by exploiting the eigenvalue decomposition of the variance-covariance matrix. Parsimonious mixtures of multivariate Gaussian distributions can be used to model random covariates within CWM-GLM or as standalone models (mixture of multivariate Gaussians with defined covariance matrix). This feature is completely new for Stata users as it is not allowed by **gsem** and **fmm**. Lastly, the flexibility of **cwmglm** allows one to estimate "canonical" finite mixture of regressions.

Stacking generalization and machine learning in Stata • Achim Arens, ETH Zurich

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 boosting, support vector machines and feed-forward neural nets (multi-layer perceptron). **pystacked** can also be used 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.

Double/debiased machine learning in Stata • Achim Arens, ETH Zurich

ddml implements algorithms for causal inference aided by supervised machine learning as proposed in "Double/debiased machine learning for treatment and structural parameters" (Econometrics Journal, 2018). Five different models are supported; allowing for binary or continuous treatment variables and endogeneity. **ddml** supports a variety of different ML programs, including but not limited to **lassopack** and **pystacked**.

13.00 - 14.00 Lunch

14.00 - 15.00 SESSION IV - EXPLOITING THE POTENTIAL OF STATA 17, II

Treatment-effects estimation using lasso • Di Liu, Senior Econometrician and Software Developer, StataCorp

One can use treatment-effects estimators to draw causal inferences from observational data. You can use **lasso** when you want to control for many potential covariates. With standard treatment-effects models, there is an intrinsic conflict between two required assumptions. The conditional independence assumption is likely to be satisfied with many variables in the model, while the overlap assumption is likely to be satisfied with fewer variables in the model. This presentation shows how to overcome this conflict by using Stata 17's **telasso** command.

telasso estimates the average treatment effects with high-dimensional controls while using **lasso** for model selection. This estimator is robust to the model-selection mistakes. Moreover, it is doubly robust, so only one of the outcome or treatment model needs to be correctly specified.

15.00 - 16.45 SESSION V - COMMUNITY CONTRIBUTED, III

Recursive bivariate probit estimation and decomposition of marginal effects • Mustafa Coban, Institute for Employment Research, Nuremberg

This article describes a new Stata command **rbiprobit** for fitting recursive bivariate probit models, which differ from bivariate probit models in allowing the first dependent variable to appear on the right-hand side of the second dependent variable. Although the estimation of model parameters does not differ from the bivariate case, the existing commands **biprobit** and **cmp** do not consider the structural model's recursive nature for post-estimation commands. **rbiprobit** estimates the model parameters, computes treatment effects of the first dependent variable and gives the marginal effects of independent variables. In addition, marginal effects can be decomposed into direct and indirect effects if covariates appear in both equations. Moreover, the post-estimation commands incorporate the two community-contributed goodness-of-fit tests **scoregof** and **bphltest**. Dependent variables of the recursive probit model may be binary, ordinal, or a mixture of both. I present and explain the **rbiprobit** command and the available post-estimation commands using data from the European Social Survey.

A Stata package to handle metadata • Gustavo Iglésias, Microdata Research Laboratory, Banco de Portugal

In this presentation, I offer a brief tour of **mdata**, a Stata user-written package that provides a set of tools to help users handle metadata in large and complex data sets. The package uses an Excel file to store all metadata related to a data set. This is particularly useful to edit and modify metadata outside of Stata, and also to deal with data sets stored in non-Stata format. The presentation will focus on the most important features of the package, namely on how to extract metadata from data in memory, perform consistency checks on the metadata, apply metadata to data in memory, compare and/or combine metadata from two data sets.

16.00 - 16.15 Coffee break

Network Regressions in Stata • Jan Ditzen, Free University of Bozen

In this talk we introduce **nwxtregress** a new community contributed routine to estimate network regressions. It uses MCMC estimation methods (LeSage and Pace 2009) 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. **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.

16.45 - 18.00 SESSION VI - APPLICATION STUDY USING STATA

Modelling the risk of multimorbidity: an application of multistate models to the Swedish National March Cohort • Giulia Peveri, University of Milan

Chronic diseases, defined as health problems requiring ongoing management over a period of years or decades [1] currently represent the predominant burden of healthcare. [2] To address the coexistence of two or more diseases or conditions, the term multimorbidity is used. When combined, chronic diseases create additional challenges to patient care, since clinical trials usually exclude patients with coexisting conditions, and therefore most guidelines do not provide recommendations for patients presenting with multiple diseases. [3] With worldwide life expectancy increasing from 45.7 years in 1950 to 72.6 years in 2019 [4] and 20% of people aged ≥ 65 years in Europe in the same year [5], understanding the patterns and risk factors of multimorbidity has become of great relevance for public health. Multi-state models are a well-suited statistical framework to address this problem.

Net Promoter Score - Beyond the Measure: a Statistical Approach Based on Generalized Ordered Logit Models Implemented by Stata to conduct a NPS Key Drivers' Analysis • Debora Giovannelli, Florence

The Net Promoter Score (NPS) index is a popular satisfaction measure which allows to gauge Customer Loyalty (CL) at most large and medium-size firms in different fields. Due to its impact on company's growth line managers are strongly interested in knowing which factors can increase NPS by increasing promoters and decreasing detractors. NPS Key Drivers' Analysis (NPS KDA) can be a suitable tool for this task. A KDA may be conducted by implementing different statistical approaches, for identifying those factors or drivers with a significant impact on a specific outcome variable. In the contest of NPS KDA, the Regression Models for ordinal outcomes represent a statistical approach for identifying those significant Customer Experience (CX) attributes which can drive Customer Status (CS) from detractors to promoters, leading companies to design appropriate improvement strategies, involving those facets of product or service with the highest improvement priority.

In this paper the NPS KDA has been conducted by implementing in Stata two special cases of the Generalized Ordered Logit Models, the Proportional Odds Model (POM)

and the Partial Proportional Odds Models (PPOM), where the dependent variable CS was modelled as function of different CX attributes.

Absences from work and climate change: an empirical analysis • Grazia Errichiello, University of Naples - Parthenope

The research aims to observe the Italian regions with most absences from work and verify if there is a relationship between the absences and climate change. Using INPS database relating to employees, the time interval taken into consideration is 2009-2018, the variable credit difference is examined, it is a measure of the salary that workers have not received due to absence from work.

Then the existence of geographical influence between Italian regions was verified through the creation of maps using the Stata software. By other variables available a new variable was created, it measures the number of absences made by workers for each region. The creation of maps made possible

to understand in which Italian regions workers are absent more. Looking only at the sectors most affected by climate change, the results vary. Finally, only sickness and injury absences were observed, as they could be caused by climate change and extreme weather events. By observing the outliner values of the variable that measures absences from work, it was found, that extreme weather events actually occurred in the month and in the region in which the value far from the average was recorded.

18.00-18.15 OPEN PANEL DISCUSSION WITH STATA DEVELOPERS • JEFF PITBLADO AND DI LIU, STATACORP

The “Open panel discussion with Stata Developers” session offers participants the opportunity to interact directly with StataCorp: providing participants with a forum in which to highlight any limitations encountered or suggest eventual improvements to the software.

20.30 Conference Social Dinner (Optional)

20 MAGGIO 2022 | 9.00-13.00 / 14.00-16.30

Massimizzare il potenziale delle nuove capacità Python di Stata

Giovanni Cerulli | IRcRES-CNR *Research Institute on Sustainable Economic Growth* | Centro Nazionale delle Ricerche

DESCRIZIONE DEL CORSO

L'integrazione di Python è una delle funzionalità più interessanti recentemente incorporate in Stata, in quanto permette agli utenti di utilizzare la vasta gamma di pacchetti Python (*open-source*) per elaborare, visualizzare ed esplorare i dati in modo interattivo all'interno dell'ambiente Stata o di incorporare codici Python direttamente nei file DO di Stata.

Il workshop "*Massimizzare il potenziale delle nuove capacità Python di Stata*" pertanto offre ai partecipanti un'eccellente opportunità per acquisire i *programming skills* necessari per integrare la capacità di Python in Stata 17 attraverso una serie di esempi che permettono di evidenziare QUANDO, e di conseguenza COME, si dovrebbe sfruttare la connettività tra Python e Stata per la propria ricerca.

L'obiettivo è di offrire una panoramica dell'applicabilità del linguaggio di programmazione Python, all'interno di Stata.

DESTINATARI: Il corso offre un'opportunità per sociologi, matematici, economisti, etnologi, epidemiologi e politologi e consente di acquisire gli strumenti di base necessari per utilizzare *routine* Python all'interno di Stata.

REQUISITI RICHIESTI: Conoscenza operativa di Stata. Non è richiesta la conoscenza di Python anche se ne comporterà un vantaggio.

PROGRAMMA

1. Panoramica sul potenziale della connettività Stata/Python
2. Nozioni di base di programmazione Python
3. Modi alternativi per implementare Python in Stata: il modulo PyStata
   - Chiamare Python da Stata
   - Chiamare Stata da Python
4. Integrazione Mata/Python
5. Esempi applicati:
   - Integrazione in Stata di Python *Scikit-learn* per il *Machine Learning*
   - Stima OLS in Mata/Python
   - Visualizzazione dei dati in Stata/Python

INFORMAZIONI GENERALI

Il materiale didattico distribuito include le dispense con la parte teorica, le *routine* (i file *do*) sviluppate per il workshop, le banche dati per l'implementazione di tutte le applicazioni empiriche e una licenza temporanea di Stata 17 valida per 30 giorni dall'inizio del workshop.

Data la natura applicata del workshop si consiglia l'utilizzo del proprio personal computer per eseguire autonomamente le sessioni applicate.

Il numero massimo di iscritti ammessi è 15 e i posti saranno assegnati in base all'ordine di arrivo.

Le richieste di partecipazione dovranno essere inviate entro il 9 Maggio 2022. Per ulteriori informazioni contattare la segreteria organizzativa a formazione@tstat.it.

TStat

www.tstattraining.eu | www.tstat.it | www.tstat.eu

LOGISTICS

The conference will be held at the Centro Studi CISL, Via della Piazzuola, 71, Florence on the 19th and 20th May 2022.

	Conference	Residential option 1 (1 night - conference)	Conference plus Training course	Residential option 2 (2 nights - conference plus training course)
Full-time Undergraduate, Masters and Ph.D Students	€ 65.00	€ 130.00	€ 245.00	€ 375.00
Others	€ 95.00	€ 160.00	€ 375.00	€ 505.00

Kindly note that:

- All participation fees are subject to VAT (applied at the current Italian rate of 22%).
- Conference fees include: coffee breaks, lunch, course materials and a temporary licence of Stata for those participants attending the training course. We have also organized a limited number of single rooms (under a Bed and Breakfast accommodation plan) at the Centro Studium CISL for those wishing also to stay in the complex.
- Individuals interested in participating in the 2022 Italian Stata Users Conference should return their duly completed registration forms to the conference organizers by the 9th May 2022 at the very latest.

FUNDING FOR DOCTORAL STUDENTS

TStat is delighted to sponsor, via our project "**Investing in Young Researchers**", THREE full-time Ph.D students from any of the countries for which TStat is the official Stata Distributor. Sponsorship covers both the first day of the conference and the training course (including accommodation on a Bed and Breakfast format). All travel expenses will however need to be paid by the participant itself.

To apply for sponsorship, please send your curriculum vitae to formazione@tstat.it .

CONTACT:

Monica Gianni - TStat | TStat Training
T. +39 0864 210101 int. 3 | formazione@tstat.it