# 2025 Northern European Stata Conference

August 29, 2025
Peter Reichard room in Biomedicum, Solnavägen 9, Karolinska Institutet

**09:00–09:30**
*wqsreg - A Stata command for Weighted Quantile Sum regression*
Marta Ponzano[1], Stefano Renzetti[2], Andrea Bellavia[3]
[1]Department of Health Sciences, University of Genoa, Genoa, Italy
[2]Department of Medicine and Surgery, University of Parma, Parma, Italy
[3]Department of Environmental Health, Harvard T.H. Chan School of Public Health Weighted Quantile Sum (WQS) regression is a flexible statistical method for quantifying the association between a set of possibly correlated predictors and a health outcome. This approach is gaining substantial popularity in several fields such as environmental epidemiology, where it allows estimating the overall effects of complex environmental mixtures as well as the specific contributions of each mixture component. A Stata command for fitting this increasingly popular procedure, however, has not been yet developed. To address this gap, we have developed a new command – wqsreg – that enables users to fit WQS regression models for continuous outcomes while allowing for the several flexible components of this framework, including: adjust for potential confounders; estimating both positive and negative overall mixture effects; providing robust weight estimates through bootstrap; specify the method used to rank variables included in the mixture (e.g. quartiles); provide iteration limits to be performed before optimization; fix the seed and customize save options. wqsreg returns the estimates from WQS regression, plots the estimated weights and creates a dataset containing the WQS index for each subject. In this talk, we will introduce the key features of WQS regression, describe wqsreg and demonstrate its use through examples. Given the increasing importance of appropriately exploring complex multidimensional exposures such as environmental mixtures, this command provides Stata users with one of the first commands to apply a modern computational approach specifically developed for these settings.

**09:30–10:00**
*Fitting joinpoint models for descriptive analysis of cancer trends in Stata*
Paul C Lambert Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway and Dept. Medical Epidemiology and Public Health, Karolinska Institutet, Stockholm, Sweden
Investigation of temporal trends of cancer incidence and mortality rates is often performed visually with interest in changes in the gradient of increases or decreases in the rates. Joinpoint models are used to help quantify the trends, using linear splines where both the number and location of the knots (joinpoints) are selected as part of the modeling process. I will describe a Stata implementation of joinpoint models and introduce the joinpoint command and associated postestimation commands. The approach can be computer intensive as all possible combinations of the number and location of knots are fitted when selecting the models. I will describe how use of Mata to fit the models leads to dramatic speed improvements. The joinpoint command has various options, for example choosing different model selection criterion and choosing the maximum number of knots and the minimum number of data points between knots. Output options include estimation of the annual percent change (APC), with two different methods to calculate confidence intervals. There is a postestimation predict command and a command to provide visual summaries of the fitted model.

**10:00–10:30**
*Stata 20 will have correct inference on random effects*
Matteo Bottai, Division of Biostatistics, IMM, KI
Mixed models, and random effects in particular, are used routinely to model data with dependent observations and effect heterogeneity. However, while random effects are convenient for specifying a model, they often complicate inference. As a result, popular software for statistical analysis often does not provide confidence intervals for random effect parameters by default, or worse, provide provably unreliable ones. This talk discusses the challenges and possible solutions.

**10:30–11:00**
Coffee break

**11:00–12:00**
*Modeling Interval-Censored Event-Time Data with Stata*
Xiao Yang, Principal Statistician and Software Developer
Do you have event-time data that you would like to model, but are unsure exactly when the events occurred? In survival analysis, interval-censored event-time data arise when the event of interest is not observed precisely but is known to have occurred within a specific time interval. Stata 17 introduced the stintcox command to fit genuine semiparametric Cox models for such data, and Stata 18 expanded its capabilities by adding support for time-varying covariates (TVCs). Building on this, Stata 19 introduces the new stmgintcox command, enabling the modeling of interval-censored multiple-event data while accounting for potential correlations between event times across different event types. In this presentation, we will describe the fundamental types of interval-censored data and demonstrate how to fit the semiparametric Cox proportional hazards model using the stintcox command. We will provide examples using single-record and multiple-record-per-subject datasets and show how to incorporate TVCs. Additionally, we will discuss how to interpret and plot results, and how to assess the proportional hazards assumption. Finally, we will show you how to fit a marginal Cox proportional hazards model to interval-censored multiple-event data and perform a more powerful test for common covariate effects across all events.

**12:00-13:00**
Lunch break

**13:00-13:30**
*Prediction Intervals in Meta-analysis: A Clearer View of Heterogeneity and Expected Future Findings Using Stata*
David J. Miller, U.S. Environmental Protection Agency (retired)
Meta-analyses in epidemiology often rely on 95% confidence intervals (CIs) to summarize the precision of pooled estimates. However, CIs are frequently misinterpreted and offer limited insight into how study results vary (heterogeneity) or what future studies might show. Prediction intervals (PIs), by contrast, directly reflect such between-study variability and estimate the range within which the true effect of a future study is expected to fall—providing a more interpretable and policy-relevant view of uncertainty. This talk presents the rationale for using PIs in meta-analyses of odds ratios (ORs), drawing on the methods described in Borenstein's widely used text on the subject. PIs will be contrasted with traditional heterogeneity measures like $I^2$, which is often misused or overinterpreted as a precise index of inconsistency. In addition, PIs allow framing heterogeneity in terms of expected future effects and provides a more intuitive and decision-relevant perspective. Using Stata, I will demonstrate how to compute and visualize PIs, including enhanced graphical methods based on probability density functions (PDFs). Such plots go beyond Stata's whisker-like default PI displays in forest plots by better illustrating both the

expected range and the relative likelihood of future effect sizes—conveying direction, dispersion, and uncertainty in a single visual. Attendees will gain a practical and conceptual understanding of how PIs can complement or even surpass CIs and $I^2$ as tools for interpreting and applying meta-analytic evidence in epidemiology.

**13:30-14:00**
*Supplementing risk ratios in sibling analysis: estimating clinically useful measures from family-based analysis*
Hugo Sjöqvist
Global Public Health, Karolinska Institutet
Family-based designs like sibling comparisons are powerful tools for addressing confounding, but they often rely solely on relative measures such as odds ratios or hazard ratios – limiting their interpretability for clinical and policy decision-making. In this talk, I introduce the marginalized between-within framework, a method that enhances family-based analyses by enabling the estimation of absolute risks and other clinically meaningful metrics. I'll begin with an overview of sibling comparison methods and the rationale behind decomposing effects into within- and between-family components. Then, using Swedish registry data, I'll demonstrate how this framework can be applied to assess the impact of maternal smoking on infant mortality. The model allows us to estimate absolute risk differences, average treatment effects, attributable fractions, and numbers needed to harm – metrics which are often more useful than relative estimates. Compared to traditional conditional logistic or stratified Cox regression models, the marginalized between-within approach offers similar relative estimates but adds the crucial ability to anchor results to a global baseline, making absolute measures possible. These measures provide clearer insights for public health and policy interventions.

**14:00-14:30**
*Imputing right-skewed bounded biomarkers in partially measured cohorts*
Nicola Orsini * Robert Thiesmeier. Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden
In large medical and epidemiological studies, important biomarkers are often only available for a limited fraction of participants due to the high laboratory costs or feasibility constraints. This results in a high proportion of missing values. Imputation strategies can be employed to prevent the loss of information. However, imputing biomarker values is challenging due to the right-skewed and naturally bounded values of biomarker distributions. In this talk, we compare two imputation strategies that can handle such challenges: a likelihood-based approach and logistic quantile imputation implemented in Stata. We evaluate the performance of both methods through simulation, assessing bias and inferential errors. The approaches are illustrated with a practical example of recently discovered blood biomarkers in Alzheimer's research. The results provide some insight on recovering biomarker distributions when outcome data are fully observed but biomarkers are only partially measured.

**14:30-15:00**
Coffee break

**15:00-16:00**
*Linking frames in Stata*
Jeff Pitblado, Executive Director, Statistical Software
This presentation gives an overview of data frames in Stata. I demonstrate the basics of working with multiple datasets in Stata. I cover most of the frames suite of commands, touching on frame creation and management, linking frames, copying variables from linked frames, alias variables, and working with a set of frames.

**16:00-17:00**
Open discussion with Stata Developers