Invited Session: Post-selection inference in regression

In regression researchers often prefer simple models. They also want to do inference on the parameters of their model. If the same data are naively used for model selection and for inference, huge biases may occur. Post-selection inference addresses the question how to do proper inference in a selected model, taking into account the fact that the same data were used twice in the analysis.

The invited speakers will be

- Ruth Heller, Department of Statistics and Operations Research, Tel-Aviv University, Israel
- Aldo Solari, Department of Economics, Management and Statistics, University of Milano-Bicocca
- Hannes Leeb, Department of Statistics and Operations Research, University of Vienna

Post-selection inference for genetic association studies

Ruth Heller, Department of Statistics and Operations Research, Tel-Aviv University

Abstract: In genetic association studies, there is a natural grouping of the genome into regions of interests (ROIs, e.g., genes), which are comprised of single variants. For powerful identification of association with the phenotype at the region level, the test statistics of single variants within an ROI can be aggregated into a test statistic for the global null hypothesis that none of the single variants are associated with the phenotype. Following ROI discovery, the focus turns to identification of the single variants that drive the association, within the ROI. Failure to account for ROI discovery (e.g., few dozen genes out of the original 20,000) can lead to biased inference. We provide post-selection inference for the family of single variants within a discovered ROI. Our inference is exact for the normal model, and asymptotically justified otherwise. We adapt our post-selection inference to the setting in which only summary level data is available from the study of interest. Analyses that only use summary level data are very attractive since they can thus avoid privacy concerns and logistics of sharing individual level data. These analyses need information about the linkage disequilibrium (LD) between the variants, which can be obtained from reference panels. We shall discuss the interesting connection between the size of the study and the reference panel used, for valid and powerful post-selection inference.

Joint work with Nilanjan Chatterjee, Tzviel Frostig, and Amit Meir

Quantifying model uncertainty by model confidence sets

Aldo Solari, Department of Economics, Management and Statistics, University of Milano-Bicocca

Abstract: A researcher often has at her disposal a collection of candidate models which could be fitted to data, and has to decide which ones are good models and which ones are bad models. Statistical inference can be used to conclude whether a model is good or bad but sometimes it leaves us with uncertainty: there is not enough evidence to reach a decision. Therefore the application of an inferential procedure provides a model confidence set, i.e. a partition of candidate models into good models, bad models and uncertain models.

On the one hand, the idea of quantifying model uncertainty by a model confidence set is interesting per se, because it provides an assessment of the power coming from the data to discriminate the models. On the other hand, a model confidence set can be used in combination with the application to the same data of model selection algorithms that deliver one or more "best" models. If the selected "best" model belongs to the set of bad models, then this selection is not admissible from the model confidence set perspective.

Two questions now arise: 1. What is the precise definition of good and bad models? 2. How to construct model confidence sets with strong inferential guarantees? We will look at these questions within the classical framework of Gaussian linear models, both with fixed and random designs.

With regard to the first question, what a good model is good for? We will first distinguish between the use of modelling for explanation and for prediction. Next, what are good and bad relative to? We will argue that from a variable selection perspective, i.e. of reducing the complexity of the full model, good and bad should be understood as relative to the full model, i.e. better than the full model and worse than the full model.

Typically model selection is not the final purpose of the analysis: for the selected models we are usually interested in e.g. estimation of their parameters or in prediction of future values. A particularly challenging task is to perform post selection inference, i.e. to perform model selection followed by statistical inference on the parameters of the selected models, all with the same data. We will discuss the relationship of the proposed approach with this challenging goal.

Statistical inference with *F*-statistics when fitting simple models to high-dimensional data

Hannes Leeb (University of Vienna and DataScience@UniVienna)
Lukas Steinberger (University of Freiburg)

Abstract: We study linear subset regression in the context of the high-dimensional overall model $y = \vartheta + \theta' z + \epsilon$ with univariate response y and a d-vector of random regressors z, independent of ϵ. Here, 'high-dimensional' means that the number d of available explanatory variables is much larger than the number n of observations. We consider simple linear sub-models where y is regressed on a set of p regressors given by $x = M'z$, for some $d \times p$ matrix M of full rank $p < n$. The corresponding simple model, i.e., $y = \alpha + \beta' x + e$, can be justified by imposing appropriate restrictions on the unknown parameter θ in the overall model; otherwise, this simple model can be grossly mis-specified. In this paper, we establish asymptotic validity of the standard *F*-test on the surrogate parameter β, in an appropriate sense, even when the simple model is mis-specified.