WORKSHOP ON CURRENT TRENDS AND CHALLENGES IN MODEL SELECTION AND RELATED AREAS

Speaker: Paul Kabaila, Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia

Title of Talk: Confidence intervals in regression utilizing prior information

Abstract:

We consider a linear regression model with independent and identically distributed zero-mean normal errors. Let p denote the dimension of the regression parameter vector and let n denote the dimension of the response vector. We suppose that the parameter of interest theta is a specified linear combination of the components of the regression parameter vector. Our aim is to find a frequentist confidence interval for theta with minimum coverage probability $1 -$ alpha. When, as a first step, a data-based model selection (e.g. by minimizing AIC or BIC or by hypothesis tests) is used to select a model, it is common statistical practice to then carry out inference about theta, using the same data, based on the (false) assumption that the selected model had been given to us *a priori*. This assumption is false and it can lead to invalid inferences.

In the first part of the talk, we consider a confidence interval for theta with nominal coverage $1 -$ alpha constructed under this (false) assumption. We call this the naive $1 -$ alpha confidence interval. We describe the methods that have been used to assess the minimum coverage probability of this confidence interval, both when $n - p$ is large and when it is not large. We also describe some of the insights that have been gained from carrying out this assessment. Typically, the minimum coverage probability of the naive confidence interval is much less than $1 -$ alpha, showing that the resulting confidence interval is completely inadequate.

Preliminary data-based model selection may be motivated by a desire to utilize uncertain prior information. In the second part of the talk, we suppose that we have uncertain prior information that the last component of the regression parameter vector is zero. Our aim is to find a frequentist confidence interval for theta that (a) has minimum coverage probability $1 -$ alpha and (b) utilizes this prior information. To utilize this prior information, this confidence interval must have expected length that (a) is relatively small when the prior information is correct and (b) has a maximum value that is not too large. We also require that this confidence interval coincide with the standard $1 -$ alpha confidence interval when the data strongly contradicts this prior information.

Consider the naive confidence interval constructed after a preliminary test of the null hypothesis that the last component of the regression parameter vector is zero against the alternative hypothesis that it is non-zero. This confidence interval typically has minimum coverage probability far below 1 – alpha. Thus this confidence interval fails abysmally to utilize the prior information.

However, we use the form of this confidence interval to motivate a new frequentist confidence interval that (a) has minimum coverage probability 1 – alpha and (b) utilizes the prior information. The performance of this new confidence interval is examined when n – p is (a) small (but greater than or equal to 1), (b) moderate and (c) large.