Katharina Parry
Massey University, New Zealand

Abstract

Title: Employing sampling techniques as variable selection for logic regression

In a standard regression problem, we have a set of variables X, whose effect on some response y is modelled. However, imagine the data looks like something straight out of the film "The Matrix" with columns of nothing but 1s and 0s, that is, the response and explanatory variables are all binary.

Furthermore, imagine we don't just want to model main effects, but are actually more interested in uncovering the effect of interactions between these binary variables on the response.

Enter logic regression – a method still in its infancy. Well, if being invented 15 years ago still counts as recent. Logic regression aims to find combinations of the explanatory variables that capture higher-order relationships in the response. It has its limits though in terms of how many explanatory variables it can handle. But as you know from the "The Matrix", sometimes there can be thousands upon thousands of such variables. But as the saying goes: do not despair, stay calm and resort to sampling. That is, we fit logic regression models to sampled subsets of explanatory variables instead. Voila. However, we need to be careful about how we sample. We want to sample in such a manner that we reduce the dimensionality without losing too much important information.

This information is stored in the hat matrix, H, a matrix that maps the responses to the fitted values in the regression analysis. We can extract the diagonal elements of H, called leverages (LS), or the off-diagonal elements, called cross-leverages (CLS). We can use both the LS and the CLS to select a sample of explanatory variables to use in our regression modelling.

In this talk, we will present some preliminary results on how well sampling works and whether we should all be despairing after all. Look forward to seeing you all there!