PROJECT TITLE: The application of machine learning techniques to model health and labour market outcomes

FIELD OF RESEARCH CODE: 1402

PROJECT SYNOPSIS:

The aim of this project is to develop statistical inference for hitherto infeasible threshold models, with application to empirical research across the social sciences. This project will focus on applications in three Australian case studies that have important policy implications for health and labour economics: (i) Body Mass Index (BMI), (ii) optimal education levels and the decriminalisation of cannabis.

i. Body Mass Index (BMI) is a widely applied metric used to correlate weight-related health status with many adverse health outcomes. Examples abound, but include: maternal and perinatal health; mental health; and disability and overall health levels. Such correlations commonly use BMI bands defined by the World Health Organization (WHO). However the arbitrary nature of such bands means that they are unlikely to
truly reflect the actual ranges that correlate with adverse health (and related) outcomes over the wide-ranging outcomes traditionally considered in this literature.

ii. The Australian workforce is characterised by having a large proportion of “overeducated” workers; those having excess amounts of education beyond that required for their job(s). Importantly, this has been shown to have many adverse effects, and has been associated with substantial earnings penalties. However, given a job-specific reference level of education, what levels of educational mismatch actually determine whether an individual is indeed overeducated or not? Moreover, do such effects vary with the extent of mismatch? Given the adverse consequences of such, clearly, this is imperative for policy-makers to understand as it will help inform educational, labour market and immigration policies alike.

iii. Decisions such as the decriminalisation of cannabis in Australia are generally made on evidence that assesses changes in behaviour (i.e., individual consumption patterns), by comparing consumption levels at two arbitrary points in time (broadly, “before” and “after”), with no relation to the specific length of time either to, or from, a specific policy change or intervention. It is possible for behaviours to adapt in anticipation of a known (or expected) future change in policy; whilst potentially quite dramatic and differential effects are likely at the time of, and immediately following, the same policy change. In the longer-term post-change, behaviours (consumption levels) may increase (or decrease), revert to pre-policy change levels, or simply follow unknown paths. Thus, in explicitly ignoring the implicit time-path(s) of behaviours pre- and post-policy change, the usual methods of evaluating policy changes are likely to result in misleading evidence and accordingly erroneous policy advice.

The above case studies represent three important, policy relevant selections from many potential applications of threshold, or breakpoint models. Indeed, there is an illustrious history, predominantly in the econometrics literature and within a linear regression-type setting, of the analysis of such, for example (also interchangeably known as regime switching and/or structural breaks, depending on the context). Invariably, though, the response variable of interest is continuous and typically linear regression techniques are used. There are significant issues with the usual approaches suggested here though. Firstly, they have been developed for linear regression models, whereas in most empirical applications in health and labour, one would typically be using nonlinear models, such as binary probits. Or, they have been developed for linear panel data models, but are only applicable in instances where the panel is balanced; the Stata command “xthreg” only works in such a context. And large-scale population level survey panel datasets, such as the well-known Household, Income and Labour Dynamics in Australia (HILDA), are ubiquitously unbalanced.

Independently, two studies have suggested an approach to circumvent these very important drawbacks, suggesting a grid-search procedure over all possible breakpoints. However, with an unknown number of breakpoints, the procedure quickly becomes infeasible. In one of the empirical examples we will consider (which has a relatively small range of the threshold variable), there are around 880 potential candidates for all possible 1-break models, but a 5-break (6-category) model would entail estimation of over $4 \times 10^{12}$ possible candidate models. Even with the recent advances in computing power, this essentially makes such an approach practically infeasible, for all but the simplest of models with a small, and known, number of breaks. However, we will avoid the testing issues that arise, for example, in the unbalanced panel data setting. This will be achieved by following recent literature and selecting the optimal model based on an information criterion. Such an approach would simultaneously determine
the number and position of all breakpoints. Relating back to the previous example, this would now only require estimation of some 4,000 models.

**FEASIBILITY AND RESOURCING – DESCRIPTION OF THE SUPPORT THIS PROJECT WILL RECEIVE:**

The student will be joining a collegial and productive group of researchers within the School of Economics, Finance and Property (SEFP), with demonstrated expertise across Health and Labour Economic analysis. They will be actively supported through access to the Faculty's HDR Hub, HDR Colloquium and training seminars designed to upskill HDR students. SEFP provides access to subscription databases relevant to this project. SEFP also provides an excellent research environment where HDRs are welcomed to participate in weekly seminars and meet with visiting academics. In short, the student will have access to all required resources supported by a collegial research environment.

**THE SIGNIFICANCE OF THE PROJECT/ PROGRAM FOR THE ENROLLING SCHOOL OR INSTITUTION:**

The proposed project:

- clearly aligns with the Faculty’s key research area of developing evidence-informed economic and social policy.
- will enhance capacity building for ECR Singh and SEFP through mentoring from experienced members of the supervision team.
- is suitable for facilitating engagement and impact activities, achieving high quality publications and developing Cat 1 grant applications.
- lends itself to significant amounts of international and collaboration.

Students must express interest in this scholarship opportunity by emailing the Project Lead listed below. Please provide a copy of your current curriculum vitae and detail your suitability to be involved in this strategic project.

**PROJECT LEAD CONTACT:**

Name: Mark Harris  
School: School of Economics, Finance and Property  
Faculty: Business and Law  
Email: Mark.harris@curtin.edu.au  
Contact Number: 08 9266 9692