| University of Cambridge Mathematics of Information (CMI)Call for External (4-6 week) Projects for First Year PhD Students |
| --- |
| organisation |
| Cardiovascular Epidemiology Unit/MRC Biostatistics Unit, University of Cambridge |
| project title |
| Instrumental variable estimation of shape of causal relationship using fractional polynomials |
| contact detals |
| External Supervisor Name: Stephen Burgess and Amy Mason **Contact email address:** [**sb452@medschl.cam.ac.uk**](mailto:sb452@medschl.cam.ac.uk)  **Contact telephone number: 01223 768259**  **Organisation address: Cambridge Institute of Public Health, Robinson Way, Cambridge, CB2 0SR** |
| BACKgrounD |
| An instrumental variable (IV) can be used to test for and estimate the causal effect of an exposure on an outcome. Mendelian randomization (MR) is an important form of instrumental variable analysis, where genes are used as the IV. MR analyses are particularly useful for assessing the potential effects of lowering the levels of a potentially causal biomarker without having to perform a randomized controlled trial. For example, research at this university has shown strong causal evidence for the effect of lipoprotein(a) on cardiovascular outcomes.  Most IV methods assume that the function relating the exposure to the expected value of the outcome (the exposure--outcome relationship) is linear. However, in practice this assumption may not hold. Indeed, often the primary question of interest is to assess the shape of this relationship. For example, we are interested in estimating the shape of the causal relationship between body mass index (BMI) and mortality. Do increases in BMI in the “normal weight” range lead to an increased risk of mortality? Is there any harm of decreasing BMI into the “underweight” category?  We have previously developed a statistical method that estimates the shape of the relationship between exposure and outcome using semiparametric methods. The paper looked at both piecewise linear models and fractional polynomials. However this method relied on having large individual level datasets – i.e. you would need the exposure value, the outcome value and the genetic make-up for each person. However most datasets only publish summary level data – the overall effect of various genes on a specific outcome. This reduces the usefulness of this approach in some cases. |
| problem |
| *Please give a description of the problem and the scope of the investigation. You may include graphs, diagrams and images. No word limit, but a half page to two pages should suffice.*  We are interested in estimating the casual effect of an exposure on an outcome. Unfortunately, the exposures and outcomes that often interest us are ones where we are fairly sure an assumption of linearity fails yet we cannot access sufficient detail for an individual level study. The ideal solution is clearly a non-linear method that works for summary level data.  We have designed such a method and written up code in R to implement it from summary data. For the individual level version of the method, a series of simulated individual data was previously created to test how accurately the models could determine the correct model to choose, and whether the true coefficients were within the 95% confidence intervals of the predicted coefficients (Staley and Burgess, 2017).  The purpose of this project would be to create similar sets of simulated data for testing the new version of the method. The student would need to create simulated summary level data with fractional polynomials as the underlying relationship between the exposure and the outcome. For example, fractional polynomials of degree 1 are defined as  where P takes values in {−2, −1, −0.5, 0, 0.5, 1, 2, 3}, and power of 0 refers to the (natural) log function. This project could concentrate solely on polynomials of this form, or could also extend to higher powers. The student would need to create summary level data that would mimic the sort of summary data available from genetic consortiums like UK Biobank.  The algorithm performance would need to be evaluated. This would be done by fitting the correct fractional polynomial (i.e. correct degree and powers) then examining bias of the parameter estimates, and how often the correct parameter is within the 95% confidence interval. |
| AIms and expectations |
| The overall aim of the project is to investigate the reliability of a fractional polynomial method when working with summary level data.  We would like the student to   * Familiarize themselves with the basic ideas behind instrumental variable estimation, Mendelian randomization and fractional polynomials. * Create simulated data for summary datasets with relationship shapes between exposures and outcomes based on 1 and 2 degree fractional polynomials * Evaluate the output of the algorithm and refine the code if needed.   We anticipate that this work would lead to a published paper. Assuming that the student makes useful progress, authorship credit would be given. The level of authorship would correspond to the commitment of the student to the project – a lead author position may require additional work beyond the 4-6 weeks of the project. |
| references and resources |
| **Video introductions:**  *An introduction to IV analysis -* <http://tiny.cc/IVvid>  *A two minute primer on MR -* <http://tiny.cc/MRvid>  *Podcast about MR -* <http://tiny.cc/MRpod>  ***General IV & MR introduction papers***  An Introduction to IV [https://www.karger.com/Article/Abstract/319455#](https://www.karger.com/Article/Abstract/319455) Bennett D, A: An Introduction to Instrumental Variables Analysis: Part 1. Neuroepidemiology 2010;35:237-240.  An introduction to MR - <https://heart.bmj.com/content/103/18/1400>  Bennett DA, Holmes MV  Mendelian randomisation in cardiovascular research: an introduction for clinicians  *Heart*2017;**103:**1400-1407.  Guide & glossary for MR - <https://www.bmj.com/content/362/bmj.k601>  Davies Neil M, Holmes Michael V, Davey Smith George. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians *BMJ*2018; 362 :k601  ***Papers specific to this project***  *Paper on using semiparametric methods on individual level data -* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5400068/>  Staley JR, Burgess S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genet Epidemiol*. 2017;41(4):341–352.  *Paper on IV on non-linear relationships*  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222800/>  Burgess S, Davies NM, Thompson SG; EPIC-InterAct Consortium. Instrumental variable analysis with a nonlinear exposure-outcome relationship. *Epidemiology*. 2014;25(6):877–885.  **No additional resources will be required by the student.** |
| requirements |
| The student will either need experience programming in R, or be confident they could quickly learn R to a working standard. We highly recommend the use of R Studio and Github for organizing keeping track of work. |
| Notes and Submission process |
| **Please return completed project proposals by 5pm on Friday 8 November 2019 to:** DrStephanie North, Knowledge Transfer Facilitator, [sn468@cam.ac.uk](mailto:sn468@cam.ac.uk) and Tessa Blackman, CMI Group Secretary, [cmi@maths.cam.ac.uk](mailto:cmi@maths.cam.ac.uk) . On completion in March 2020, the students will write a short report and give a brief presentation on their project to the cohort of students, academic and industrial supervisors. In addition, we require feedback from the external supervisors on the outcome and any impact resulting from the project.  **Thank you for participating.** |

Plan

1. Get function working
2. Individual level data set (see original paper)
3. Use to create summary level data set
4. Compare answers
5. Repeat with a practical analysis (Biobank?)

Another thing that would be nice to show is that the summary data analysis gives the same answer as the individual-level data analysis. Could do this for an applied example from UK Biobank – perform the analysis using summarized and individual-level data.

When you’ve put some thought into the methods, would be good to sketch out a plan of a potential paper – ie what the simulation analysis would look like, what applied example to include.