Graduate School Workshop on Big Data

Professor John Bacon-Shone, Director, Social Sciences Research Centre

Aims

On completion, students should know how to handle building statistical models for datasets which are too large to fit into memory on a typical computer.

Organization

Students must have their own laptop (Windows, Mac or Linux are all fine), but do not need to pre-install any software as installation will be part of the workshop.

The workshop will be run using Zoom, so participants need to be able to run Zoom on that day in order to connect remotely, in order to benefit from the interactive teaching. The session will be recorded to benefit those unable to join us live.

Our focus will be on the dataset below, but we will spend some time dividing up the work and then discussing the different answers we get from different approaches.

Assumed knowledge

Students will be assumed to already understand linear models well. I will provide a downloadable textbook in advance for those who need revision. Knowledge of using R is beneficial.

References

The dataset we will explore can be found here:
http://stat-computing.org/dataexpo/2009/the-data.html

It contains about 120 million records on 29 core variables.