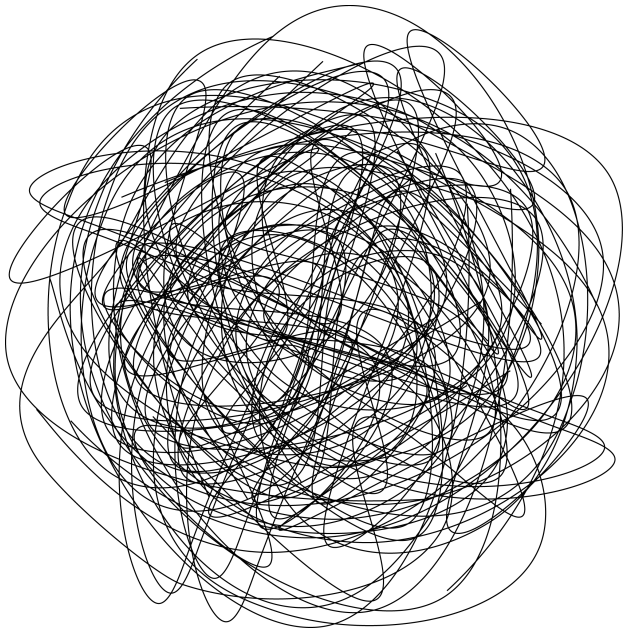
THE UNIVERSITY *of* EDINBURGH
Centre for Statistics

Dealing with Messy Data

March 8, Wednesday 2023
Nucleus Building, Larch Lecture Theatre ($NUC_{1.15}$)
University of Edinburgh, King's Buildings, EH9 3FD



Hosted by the **Centre for Statistics** and **School of Social and Political Science**
https://centreforstatistics.maths.ed.ac.uk
https://www.sps.ed.ac.uk/

The **'Dealing with Messy Data' event** event is less about showcasing finalised projects and more about presenting an ongoing challenge to be discussed with the CfS group. It is hoped that joint discussion can potentially lead to collaboratively sought solutions to analytical challenges. Statisticians may wish to present methodology that may be relevant or interesting to the applied researchers who are tackling analytical challenges brought about by complex and imperfect data available for analysis.

PROGRAMME:

- 13:30 - 13:45: Welcome & Coffee
- 13:45 - 14:15: Dr Roxanne Connelly (Senior Lecturer of Sociology and Quantitative Methods)
  The challenge of calculating model fit statistics for logistic regression models using complex survey data
- 14:15 - 14:45: Dr Orian Brook (Chancellor's Fellow in Social Policy)
  Brainstorming solutions to address missing data in the ONS Longitudinal Study
- 14:45 - 15:15: Coffee Break.
- 15:15 - 15:45: Dr Christopher Barrie (Lecturer in Computational Sociology)
  Machine Learning Models for the Measurement of Media Criticism
- 15:45 - 16:15: Dr Ruth King (Thomas Bayes' Chair of Statistics)
  Multiple Systems Estimation: An Approach for Estimating Difficult to Observe Populations
- 16:15 – 16:30: Closing remarks and spill-over time (+ Group Photo).

Invited talks are designed to be 15 min with 15 min Q&A and changeover. The full programme including titles and abstracts of talks can be on the CfS website.

The challenge of calculating model fit statistics for logistic regression models using complex survey data

Roxanne Connelly

Senior Lecturer of Sociology and Quantitative Methods

Abstract: Nationally representative social survey data resources almost always use complex designs which incorporate clustering, stratification and unequal selection probabilities (e.g. the UK Household Longitudinal Study, the Millennium Cohort Study). Statistical data analysis packages (e.g. Stata and R) now have routines for the analysis of many mainstream statistical models. However, a challenge remains when calculating model fit statistics for models of categorical outcome variables (e.g. logit).

†**Time slot**: 13:45-14:15

Brainstorming solutions to address missing data in the ONS Longitudinal Study

Orian Brook

Chancellor's Fellow in Social Policy

Abstract: In research undertaken by Dr Brook which used the Office of National Statistics Longitudinal Study, creative ways to deal with and address missing data across four censuses had to be devised. This presentation discusses one approach taken to address missing observations in this specific setting, and invites suggestions regarding what else can be done in this and similar research settings.

†**Time slot**: 14:15-14:45

Machine Learning Models for the Measurement of Media Criticism

Christopher Barrie

Lecturer in Computational Sociology

Abstract: The ability of news media to criticize government is a core pillar of media freedom. Existing indices tend to use scoring criteria or expert surveys to develop over-time measures of media freedom. In this article, we use the largest existing dataset of Arabic-language news to evaluate how political reporting changes over the course of a successful and failed democratic transition. Using entirely unsupervised ALC word-embedding techniques, we demonstrate how to generate temporally granular measurements of media criticism that closely correlate with measurements derived from expert surveys for both Egypt and Tunisia. Crucially, the technique we propose is computationally inexpensive, effectively cost-free, and eminently scalable. In this, our work points to new possibilities in the monitoring and measurement of media capture within authoritarian and transitional settings.

†**Time slot**: 15:15-15:45

Multiple Systems Estimation: An Approach for Estimating Difficult to Observe Populations

Ruth King

Thomas Bayes' Chair of Statistics

Abstract: In this talk I will give a brief summary of multiple systems estimation, often used in the estimation of hidden or difficult to observe populations. Applications, include, for example, the number of people who inject drugs; homeless individuals in a given area; civilian casualties in a war, etc. The approach uses a combination of multiple administrative data lists, where individuals are uniquely identifiable by each data list. This permits the cross-classification of individuals across the lists, with the data typically summarised as the number of individuals observed by each distinct combination of lists. These data are subsequently used to obtain an estimate of the number of individuals not observed by any list (called the hidden or dark figure), which when combined with the number of observed individuals is an estimate of the total population size. In particular, generalised linear models (GLM) are applied that are able to account for potential interactions between the different lists (e.g. being on one list may make it more/less likely of being observed on another list).

A number of issues arise and need to be addressed when applying a multiple systems estimation approach. These include, for example, model selection (in terms of the interactions present in the model) and associated model averaging approach to incorporate model uncertainty; incorporating additional external prior information; including covariate information (such as gender, age) and dealing with "censored" cells. I will briefly discuss a number of these issues and some current ongoing research.

†**Time slot**: 15:45-16:15

Organized by

Natalia Bochkina (n.bochkina@ed.ac.uk), Valeria Skafida (valeria.skafida@ed.ac.uk), Roxanne Connelly (roxanne.connelly@ed.ac.uk), Amanda Lenzi (amanda.lenzi@ed.ac.uk), Torben Sell (torben.sell@ed.ac.uk), Ozan Evkaya (oevkaya@ed.ac.uk).