Pre-Analysis Plan

The Stark Realities of Reproducible Sociological Research

Professor Vernon Gayle, University of Edinburgh, UK
vernon.gayle@ed.ac.uk

9th June 2020

Draft 1.0

Contents

Authorship and Meta Information

Author: Professor Vernon Gayle

Orcid id: http://orcid.org/0000-0002-1929-5983

Project: Reproducible Sociological Research

Background

Across a wide range of academic disciplines there is increasing concern that research findings cannot be reproduced (i.e. consistently repeated), and therefore it is impossible to verify empirical results (Yale, 2010, Nature, 2016, Baker, 2016, Christensen et al., 2019). A number of reproducibility guidelines have been proposed. Baiocchi (2007) proposed guidelines for computational economics, Hofner et al. (2016) for biometrics, Begley and Ioannidis (2015) for medical research, Sandve et al. (2013) for computational research, and Nosek et al. (2012) for psychology. At the current time there is little guidance for sociological researchers.

Research Question

Can a sociological researcher follow Professor Philip Stark's checklist for reproducible research (see http://www.bitss.org/2015/12/31/science-is-show-me-not-trust-me/ [1]) and undertake a plausible piece of analysis, using genuine large-scale data?

[1] Web paged archived using https://archive.org/web/ on 9th June 2020.

Data Analysis Software

The data enabling and data analysis in this work will principally be undertaken in *R*. The decision to use *R* is motivated by an attempt to use an open source data analytical software. Other data analysis software such as Stata, SPSS and Python might also be required.

The Research Datasets

The Youth Cohort Study of England and Wales (YCS)

The Youth Cohort Study of England and Wales (YCS) is a major longitudinal study that began in the mid-1980s. It was a large-scale nationally representative survey funded by the government and was designed to monitor the behaviour of young people as they reached the minimum school leaving age and either remained in education or entered the labour market.

There are a number of challenges associated with analysing YCS data, most notably inadequate documentation on the procedures used to construct the datasets.

1. YCS Cohort Nine (1998-2000) UK Data Archive Study 4009

https://discover.ukdataservice.ac.uk/catalogue/?sn=4009

The population studied were male and female school pupils in England and Wales who had reached minimum school leaving age in the 1996/1997 school year. To be eligible for inclusion they had to be aged 16 on August 31st 1997.

Downloaded: UK Data Service https://www.ukdataservice.ac.uk/

Date: 8th June 2020

Time: 15:14

Russell, N., Finch, S., McAleese, I., Nice, D., Fitzgerald, R., La Valle, I. (2004). *Youth Cohort Study of England and Wales, 1998-2000; Cohort Nine, Sweep One to Four*. [data collection]. *5th Edition.* UK Data Service. SN: 4009, http://doi.org/10.5255/UKDA-SN-4009-1

2. *Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 UK Data Archive Study 5765*

https://discover.ukdataservice.ac.uk/catalogue/?sn=5765

The Education and Youth Transitions project (EYT) was funded by the ESRC from 2003 to 2006.

A key part of the project was to create comparable time-series datasets for England, Wales and Scotland from the Youth Cohort Study (YCS) and Scottish School Leavers Survey (SSLS).

Downloaded: UK Data Service https://www.ukdataservice.ac.uk/

Date: 8th June 2020

Time: 15:14

Shapira, M., Iannelli, C., Croxford, L. (2007). *Youth Cohort Time Series for England, Wales and Scotland, 1984-2002*. [data collection]. Scottish Centre for Social Research, University of Edinburgh, Centre for Educational Sociology, National Centre for Social Research, [original data producer(s)]. Scottish Centre for Social Research. SN: 5765, http://doi.org/10.5255/UKDA-SN-5765-1

Potential Variables

YCS Cohort Nine (1998-2000) UK Data Archive Study 4009

"serial" "weight" "sex"
"s1a_c" "a58" "s1eth"
"s1acqe" "pseg" "pseg1"
"pseg2" "pseg3" "pseg4"
"pseg5" "pseg6" "pseg7"

Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 UK Data Archive Study 5765

"t0cohort" "t0nation" "t0caseid"
"t0source" "t1weight" "t1resp"
"t0parsec"

Data Analysis

1. Duplication of the Logistic Regression Model Reported in Connolly (2006)

Model reported in Table 5 p.20 Connolly (2006).

Logistic regression model of 5+ GCSEs at Grades A* - C (0,1).

Explanatory Variables: Gender, Ethnicity, Social Class.

2. Replication of the Logistic Regression Model Reported in Connolly (2006) Adding Quasi-Variance Based Estimates

Model reported in Table 5 p.20 Connolly (2006) with the addition of quasi-variance based estimates (see Gayle and Lambert, 2007).

3. Replication of Logistic Regression Model Reported in Connolly (2006) Adding National Socio-economic Classification (NS-SEC) Measure Social Class from UK Data Archive Study 5765

National Socio-economic Classification (NS-SEC) (see Rose and Pevalin, 2003).

Deliverables

1. A reproducible workflow within a Jupyter notebook
2. A data dictionary (codebook) accompanying the work
3. A ‘research object’ deposited in an open repository

References

BAIOCCHI, G. 2007. Reproducible research in computational economics: guidelines, integrated approaches, and open source software. *Computational Economics,* 30**,** 19-40.

BAKER, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature,* 533**,** 452-454.

BEGLEY, C. G. & IOANNIDIS, J. P. 2015. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research,* 116**,** 116-126.

CHRISTENSEN, G., FREESE, J. & MIGUEL, E. 2019. *Transparent and reproducible social science research: How to do open science*, University of California Press.

CONNOLLY, P. 2006. The effects of social class and ethnicity on gender differences in GCSE attainment: a secondary analysis of the Youth Cohort Study of England and Wales 1997–2001. *British Educational Research Journal,* 32**,** 3-21.

GAYLE, V. & LAMBERT, P. 2007. Using Quasi-Variance To Communicate Sociological Results From Statistical Models. *Sociology,* 41**,** 1191-1208.

HOFNER, B., SCHMID, M. & EDLER, L. 2016. Reproducible research in statistics: A review and guidelines for the Biometrical Journal. *Biometrical journal,* 58**,** 416-427.

NATURE 2016. Reality check on reproducibility. *Nature,* 533.

NOSEK, B. A., SPIES, J. R. & MOTYL, M. 2012. Scientific Utopia:II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science,* 7**,** 615-631.

ROSE, D. & PEVALIN, D. J. 2003. *A researcher's guide to the National Statistics Socio-economic Classification,* London, SAGE Publications.

SANDVE, G. K., NEKRUTENKO, A., TAYLOR, J. & HOVIG, E. 2013. Ten simple rules for reproducible computational research. *PLoS computational biology,* 9**,** e1003285.

YALE 2010. Law School Roundtable on Data and Code Sharing Reproducible Research. *Computing in Science & Engineering* 12**,** 8–13.