Pre-Analysis Plan

**The Stark Realities of Reproducible Research**

Professor Vernon Gayle, University of Edinburgh, UK

[vernon.gayle@ed.ac.uk](mailto:vernon.gayle@ed.ac.uk)

24th June 2017

Draft 1.0

**Table of Contents**

**Authorship and Meta Information2**

**Research Question 2**

**Literate Computing2**

**Data Analysis Software3**

**The Research Datasets**

**The Youth Cohort Study of England and Wales (YCS)** **3**

**Variables4**

**Data Analysis**

**Duplication of Logistic Regression Model Reported in Connolly (2006) 5**

**Replication of Logistic Regression Model Reported in Connolly (2006)**

**Using Quasi-Variance based Estimation5**

**Replication of Logistic Regression Model Reported in Connolly (2006)**

**Adding National Socio-economic Classification (NS-SEC) Measure Social Class from UK Data Archive Study 57656**

**Deliverables6**

**Authorship and Meta Information**

Author: Professor Vernon Gayle

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

Project: Reproducible Sociological Research

**Research Question**[¶](http://localhost:8888/notebooks/noobs_20170624_vg_v10.ipynb#Research-Question-)

Can a sociological researcher follow Professor Philip Stark's checklist for reproducible research and undertake a plausible piece of analysis, using genuine large-scale data with realistic levels of messiness?

See <http://www.bitss.org/2015/12/31/science-is-show-me-not-trust-me/> .

## **Literate Computing**[¶](http://localhost:8888/notebooks/noobs_20170624_vg_v10.ipynb#Literate-Computing)

Fernando Perez states that

"Literate Computing is the weaving of a narrative directly into a live computation, interleaving text with code and results to construct a complete piece that relies equally on the textual explanations and the computational components, for the goals of communicating results in scientific computing and data analysis" (see <http://blog.fperez.org/>).

Literate programming is a paradigm introduced by Donald Knuth (<http://cs.stanford.edu/~uno/>) in which a program is given as an explanation of its logic in a human readable language (e.g. plain English) with snippets traditional source code (or macros) (see <https://en.wikipedia.org/wiki/Literate_programming>).

A challenge of this current project is simple - **can I undertake a plausible piece of analysis, using genuine large-scale data with realistic levels of messiness, that is 'literate' as well as reproducible?**

**Data Analysis Software**

The data enabling and data analysis that will be undertaken in this paper will mainly be undertaken in *R*. The decision to use *R* is motivated by an attempt to use and open source data analytical software rather than a proprietary software package. Other data analysis software such as Stata, SPSS and Python might also be used.

**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 is a large-scale nationally representative survey funded by the government and is designed to monitor the behaviour of young people as they reach the minimum school leaving age and either remain in education or enter the labour market.

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

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

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

The population studied was 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: 19th June 2017

Time: 19:54

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

1. ***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: 23rd June 2017

Time: 00:17

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

**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 Logistic Regression Model Reported in Connolly (2006)***

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

Explanatory Variables: Gender, Ethnicity, Social Class.

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

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*(1), pp.3-21

1. ***Replication of Logistic Regression Model Reported in Connolly (2006)***

***Using Quasi-Variance based Estimation***

For a discussion of this method see Gayle, V. and Lambert, P.S., 2007. Using quasi-variance to communicate sociological results from statistical models. *Sociology*, *41*(6), pp.1191-1208.

1. ***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***

For a discussion of this measure see Rose, D. and Pevalin, D.J., 2002. *A researcher's guide to the national statistics socio-economic classification*. Sage.

***Deliverables***

1. A reproducible workflow within a Jupyter notebook deposited in a Git repository
2. A data dictionary (codebook) accompanying the work