The Stark Realities of Reproducible Sociological Research

Alternative titles: Some Newer Rules of the Sociological Method or The Moon Under Water

Professor Vernon Gayle, University of Edinburgh, UK.

Please remember that this work is very exploratory.

Positive comments are always appreciated, but brickbats improve work. Here is how to contact me or @profbigvern.

Next Actions:

1. Share this notebook with colleagues.

Table of Contents:

Notebook Meta Information

Computing Environment and Software

Research Meta Information

Data Description

Data Enabling

Data Analysis

Data Documentation

Commentary

Notes

License

This work must not be copied or cited without written permision from the author

When the work is nearer completion I will make it more open with an

The MIT License (MIT)

which will say...

Authorship and Meta Information

Author: Professor Vernon Gayle Orcid id: http://orcid.org/0000-0002-1929-5983

Project: Reproducible Sociological Research

Sub-project: Stratification Conference (Edinburgh) September 2017

Using this Notebook

Using Jupyter notebooks for large-scale social science data analysis in sociology is zygotic.

This is an early example of undertaking a complete analytical workflow within a Jupyter notebook.

As this practice becomes more ubiquitos it is likely that there will be improvements and best practices will become much more evident.

Warning.

Within this Jupyter notebook there has been a lot of non-routine work. For example I have 'swivel-chaired' between data analytical software packages and changed kernels.

It may from time to time be necessary to re-start the notebook depending on how stable your computing environment is.

Therefore in some sections I re-start a *R* session.

Please remember that this work is very exploratory.

Positive comments are always appreciated, but brickbats improve work. Here is how to contact me or @profbigvern.

Updates

Latest Update: 28th June (pushed up to https://github.com/vernongayle/new_rules_of_the_sociological_method) and e-mailed to colleagues.

Previous Updates:

27th June (my Mum's birthday) General work 26th June Ethics Approval Form Submitted 25th June Pre-paratory work begins 24th June Pre-Analysis Plan submitted to date stamping

---

Pre-Analysis Plan

A pre-analysis plan is openly available (in word format) https://github.com/vernongayle/new_rules_of_the_sociological_method/blob/master/pre_analysis_plan_
.

The pre-analysis plan has been formally timestamped by Originstamp.

hash: ca0fc7d948fd67cf8a1a2ac9111e9bf40425c010dfdf76ef33a0e578a90981a8

Submitted to OriginStamp: 24 Jun 2017 21:00:24 GMT Submitted to the Blockchain: 25 Jun 2017 16:00:21 GMT

This document can be verifyied using the hash at https://app.originstamp.org/verify .

Note: Some researcher might not consider this document to be a pure pre-analysis plan. This is because I have examined the data and worked with it previously. However, it is an example of how a pre-analysis plan could work in the sociological analysis of an existing large-scale social survey dataset.

Research Ethics Approval Application

A research ethics approval application has been made to the School of Social and Political Science, University of Edinburgh https://github.com/vernongayle/new_rules_of_the_sociological_method/blob/master/Research_Ethics_ .

Research Ethics Approval

From: MOORE Niamh Sent: 26 June 2017 17:01 To: GAYLE Vernon Cc: SSPS Research Subject: FW: Ethics form submission (Vernon Gayle: The Stark Realities of Reproducible Sociological Research)

Hi Vernon,

Approved at level 1. If only they were all so straightforward.

Good luck with the project.

All the best

Niamh

All the best with your application.

Niamh

Dr Niamh Moore

Chancellor's Fellow | Deputy Director of Research (Ethics) Sociology | Room 3.09, 3F2 | 18 Buccleuch Place
School of Social and Political Sciences | University of Edinburgh | Edinburgh EH8 8LN

niamh.moore@ed.ac.uk | @rawfeminism | +44(0)131-6508260 | skype: niamhresearcher
http://www.sociology.ed.ac.uk/people/staff/niamh_moore

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?

Overview of the Reproducibility Checklist

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

Philip Stark outlines 14 reproducibility points that an analysis can fail on

1. If you relied on Microsoft Excel for computations, fail.
2. If you did not script your analysis, including data cleaning and munging, fail.
3. If you did not document your code so that others can read and understand it, fail.
4. If you did not record and report the versions of the software you used (including library dependencies), fail.
5. If you did not write tests for your code, fail.
6. If you did not check the code coverage of your tests, fail.
7. If you used proprietary software that does not have an open-source equivalent without a really good reason, fail.
8. If you did not report all the analyses you tried (transformations, tests, selections of variables, models, etc.) before arriving at the one you chose to emphasize, fail.
9. If you did not make your code (including tests) available, fail.
10. If you did not make your data available (and a law like FERPA or HIPPA doesn’t prevent it), fail.
11. If you did not record and report the data format, fail.
12. If there is no open source tool for reading data in that format, fail.
13. If you did not provide an adequate data dictionary, fail.
14. If you published in a journal with a paywall and no open-access policy, fail.

Literate Computing

Fernando Perez says

"*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 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 sub-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?**

Computing Environment and Software

Computing Environment

Work undertaken using machine surface pro 109.152.252.166 (my public IP address).

Processor: Intel(R) Core™ i5-4300U CPU@ 1.90 GHz 2.50 GHz
Installed memory (RAM): 4.00 GB
System type: 64-bit Operating System, x-64 based processor

Data Analysis Software

R Analysis

The data analysis that will be undertaken in this paper will mainly be undertaken in *R*.

The decision to use *R* is motivated by checklist item 7, and it is an attempt to use and open source data analytical software rather than a proprietary software package.

WARNING

You must have *R* installed on your machine.

You must have installed the *R* kernel (See https://anaconda.org/r/r-irkernel).

You must have installed the *R* libraries foreign and survey

for example run the code *install.packages("foreign","survey")*

(see https://cran.r-project.org/web/packages/survey/index.html ; https://cran.r-project.org/web/packages/foreign/index.html).

Reminder *Switch Kernel to R < Menu kernel - change kernel>*

Getting the libraries for *R*

In [1]:

```r
library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package: grid
Loading required package: Matrix
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
```

```
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns
```

Various WARNINGS will appear. Don't panic.

If you have a more serious ERROR message here it is possibly because you have not switched to the _R Kernel_.

The Code Test

Part 1 Logistic Regression

In this block of the work I undertake a test of the software.

Because the analyses below will be based on a logistic regression model I have chosen to check the results of a logit model in my software environment against a known result.

In this section I will import a dataset from the Stata website (www.stata-press.com) and estimate a logit model.

In [3]:

```
myautodata <- read.dta("http://www.stata-press.com/data/r12/auto.dta")
```

In [4]:

```
summary(myautodata)
```

Out[4]:

```
     make               price            mpg             rep78      
 Length:74          Min.   : 3291   Min.   :12.00   Min.   :1.000  
 Class :character   1st Qu.: 4220   1st Qu.:18.00   1st Qu.:3.000  
 Mode  :character   Median : 5006   Median :20.00   Median :3.000  
                    Mean   : 6165   Mean   :21.30   Mean   :3.406  
                    3rd Qu.: 6332   3rd Qu.:24.75   3rd Qu.:4.000  
                    Max.   :15906   Max.   :41.00   Max.   :5.000  
                                                    NA's   :5      
    headroom          trunk           weight         length           turn      

 Min.   :1.500   Min.   : 5.00   Min.   :1760   Min.   :142.0   Min.   :31.0
0  
 1st Qu.:2.500   1st Qu.:10.25   1st Qu.:2250   1st Qu.:170.0   1st Qu.:36.
00  
 Median :3.000   Median :14.00   Median :3190   Median :192.5   Median :40.
00  
 Mean   :2.993   Mean   :13.76   Mean   :3019   Mean   :187.9   Mean   :39.6
5  
 3rd Qu.:3.500   3rd Qu.:16.75   3rd Qu.:3600   3rd Qu.:203.8   3rd Qu.:43.
00  
 Max.   :5.000   Max.   :23.00   Max.   :4840   Max.   :233.0   Max.   :51.0
0  
                                                                                
  displacement     gear_ratio        foreign  
 Min.   : 79.0   Min.   :2.190   Domestic:52  
 1st Qu.:119.0   1st Qu.:2.730   Foreign :22  
 Median :196.0   Median :2.955                
 Mean   :197.3   Mean   :3.015                
 3rd Qu.:245.2   3rd Qu.:3.353                
 Max.   :425.0   Max.   :3.890                
```

Estimating a logistic regression model.

- Outcome variable = *foreign*
- Explanatory variables = *weight* + *mpg*

In [7]:

```
myautologit1 <- glm(foreign ~ weight + mpg,  data = myautodata, family = "b
inomial")
```

Summarizing the output of the logistic regression model.

In [8]

In [8]:

```
summary(myautologit1)
```

Out[8]:

```
Call:
glm(formula = foreign ~ weight + mpg, family = "binomial", data = myautodat
a)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0436  -0.4285  -0.2207   0.5347   2.0679  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) 13.708367   4.518709   3.034 0.002416 ** 
weight      -0.003907   0.001012  -3.862 0.000113 ***
mpg         -0.168587   0.091917  -1.834 0.066637 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 90.066  on 73  degrees of freedom
Residual deviance: 54.350  on 71  degrees of freedom
AIC: 60.35

Number of Fisher Scoring iterations: 6
```

These results are identical to the results that are found in the Stata Manual p.1271.

Therefore I am confident that the software environment is providing the correct results for a logistic regression model.

Because I intend to use quasi-variances I will also test analyses below will be based on a logistic regression model I have chosen to check the results of a logit model in my software environment against a known result.

In this section I will import a dataset from the Stata website (www.stata-press.com) and estimate a logit model.

Part II Quasi-Variance Estimation

In the replication part of the analysis I intend to calculate quasi-variance estimates after estimating the logistic regression model. As a furth code test I will use the ship damage data and estimate an overdispersed poisson loglinear model for ship damage data from McCullagh and Nelder (1989), Sec 6.3.2.

Make sure that in R qvcalc is installed

code required in *R*
install.packages('qvcalc')

```
In [22]:
```

```r
library(MASS)
library(qvcalc)
data(ships)
ships$year <- as.factor(ships$year)
ships$period <- as.factor(ships$period)
shipmodel <- glm(formula = incidents ~ type + year + period,
    family = quasipoisson,
    data = ships, subset = (service > 0), offset = log(service))
summary(shipmodel)
shiptype.qvs <- qvcalc(shipmodel, "type")
summary(shiptype.qvs, digits = 4)
plot(shiptype.qvs)
```

```
Warning message:
: package 'qvcalc' was built under R version 3.2.5
```

```
Out[22]:

Call:
glm(formula = incidents ~ type + year + period, family = quasipoisson,
    data = ships, subset = (service > 0), offset = log(service))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6768  -0.8293  -0.4370   0.5058   2.7912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.40590    0.28276 -22.655  < 2e-16 ***
typeB       -0.54334    0.23094  -2.353  0.02681 *
typeC       -0.68740    0.42789  -1.607  0.12072
typeD       -0.07596    0.37787  -0.201  0.84230
typeE        0.32558    0.30674   1.061  0.29864
year65       0.69714    0.19459   3.583  0.00143 **
year70       0.81843    0.22077   3.707  0.00105 **
year75       0.45343    0.30321   1.495  0.14733
period75     0.38447    0.15380   2.500  0.01935 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.691028)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

Model call:  glm(formula = incidents ~ type + year + period, family = quasi
poisson,      data = ships, subset = (service > 0), offset = log(service))
Factor name:  type
      estimate     SE quasiSE quasiVar
    A  0.00000 0.0000  0.2010  0.04039
    B -0.54334 0.2309  0.1127  0.01270
    C -0.68740 0.4279  0.3753  0.14081
    D -0.07596 0.3779  0.3239  0.10491
    E  0.32558 0.3067  0.2322  0.05390
    E  0.32558 0.3067  0.2322  0.05390
Worst relative errors in SEs of simple contrasts (%):  -0.7 0.9
Worst relative errors over *all* contrasts (%):  -2.1 1.6
```

These results are identical to the results that are found in Firth, D. and De Menezes, R.X., 2004. Quasi-variances. Biometrika, pp.65-80.

Therefore I am confident that the software environment is providing the correct results for quasi-variance estimates.

The Research Dataset

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.

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

Data Enabling (a real attempt with the raw data)

Description of the dataset

The dataset used in this set of analyses is from YCS cohort 9 - sweep 1.

The file is called "ycs9sw1".

The file will be read in Stata format (i.e. th ycs9sw1.dta).

Data Wrangling (a real attempt with the raw data)

In [1]:

```
# If you have not run the notebook sequentially...

# theses libraries are required

library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package:
grid
Loading required package: Matrix
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
```

```
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns

```

Various WARNINGS will appear. Don't panic!

This file is located on (my) OneDrive.

An internet connection is useful but it is not required, as the dataset is stored locally.

When reading your version of the data from your specific location....
Remember that C:/temp/ (is windows) is C:/data/ here in the Jupyter notebook.

In [10]:

```
# This file is located on (my) OneDrive.

mydata.df <- read.dta("C:/Users/Vernon/OneDrive - University of Edinburgh/D
ocuments/ycs_9_2017/UKDA-4009-stata8/stata8/ycs9sw1.dta")
```

```
Warning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecated
```

Various WARNINGS will appear. Don't panic!

These error messages occur because *R* is reading a *Stata* .dta file. It is a genuine large-scale research dataset and includes a large number of value labels and variable labels.

In [3]:

```
summary(mydata.df)
```

```
Warning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
```

```
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecatedWarning message:
In `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels) else past
e0(labels, : duplicated levels in factors are deprecated
```

Out[3]:

```
     serial            weight                 s1estab                  s1gor      
 Min.   :200001   Min.   :0.6025   lea comp 18:6207   south east    :2365  
 1st Qu.:206123   1st Qu.:0.7661   lea comp 16:4179   eastern       :1608  
 Median :211589   Median :0.8779   gm  comp 18:1617   west midlands:1595  
 Mean   :212056   Mean   :1.0000   independent:1053   london        :1572  
 3rd Qu.:217027   3rd Qu.:1.0576   gm  comp 16: 536   north west    :1506  
 Max.   :231392   Max.   :2.5176   modern      : 454   south west    :1423  
                                    (Other)     : 616   (Other)       :4593  
                   sex      
 not answered (9)    :   0  
 item not applicable:   0  
 male                :6889  
 female              :7773  
```

```
                                               s1resp
not answered (99)                                 :   0
item not applicable                               :   0
postal mailout 1                                  :9311
postal mailout 3                                  :2564
postal mailout 4                                  :1299
questionnaire sent in response to telephone chaser:  34
telephone interview                               :1454
                 a1_a                        a1_b
not answered (9)   :    0   not answered (9)   :   0
item not applicable:    0   item not applicable:   0
agree              :10065   agree              :4787
disagree           : 4166   disagree           :9369
NA's               :  431   NA's               : 506


                 a1_c                       a1_d
not answered (9)   :   0   not answered (9)   :    0
item not applicable:   0   item not applicable:    0
agree              :9875   agree              :12124
disagree           :4267   disagree           : 2074
NA's               : 520   NA's               :  464


                 a2_1                        a2_2
not answered (9)   :    0   not answered (9)   :    0
item not applicable:    0   item not applicable:    0
yes                :13837   yes                :13393
no                 :  732   no                 :  413
NA's               :   93   NA's               :  856


                 a2_3                       a3_1
not answered (9)   :   0   not answered (9)   :    0
item not applicable:   0   item not applicable:    0
a great deal       :2714   yes                :12868
quite a lot        :7236   no                 : 1413
not much           :3018   not sure           :  289
nothing at all     : 405   NA's               :   92
NA's               :1289
                a3_1a                      a3_1b
not answered (9)   :   0   not answered (9)   :   0
item not applicable:   0   item not applicable:   0
yes                :9193   yes                :6325
no                 :3623   no                 :6448
NA's               :1846   NA's               :1889


                 a4_1                        a4_2
not answered (9)   :    0   not answered (9)   :    0
item not applicable:    0   item not applicable:    0
yes                :13961   yes                :13305
no                 :  623   no                 :  633
NA's               :   78   NA's               :  724


                a4_3a                       a4_3b
not answered (9)   :    0   not answered (9)   :    0
item not applicable:    0   item not applicable:    0
yes                :11147   yes                :11131
```

```
no                 : 2059   no                 : 2103
NA's               : 1456   NA's               : 1428


                a4_3c                       a5_1
not answered (9)   :   0   not answered (9)   :   0
item not applicable:   0   item not applicable:   0
yes                :6281   yes                :8829
no                 :6892   no                 :5036
NA's               :1489   NA's               : 797


                a5_2a                      a5_2b
not answered (9)   :   0   not answered (9)   :   0
item not applicable:   0   item not applicable:   0
yes                :7244   yes                :7152
no                 :1531   no                 :1611
NA's               :5887   NA's               :5899


                a5_2c                      a6_1a
not answered (9)   :   0   not answered (9)   :    0
item not applicable:   0   item not applicable:    0
yes                :3606   yes                :12548
no                 :5117   no                 : 1969
NA's               :5939   NA's               :  145


                a6_1b                        a7
not answered (9)   :   0   not answered (9)   :   0
item not applicable:   0   item not applicable:   0
very               : 888   yes                :9318
fairly             :6904   to some extent     :3975
not very           :3856   no                 :1257
not all            : 879   NA's               : 112
NA's               :2135
                  a8
not answered (9)   :   0
item not applicable:   0
very easy          :2684
fairly easy        :7730
fairly difficult   :3380
very difficult     : 698
NA's               : 170
                                                          a9_1
full-time education at school or a college of further educat:10901
modern apprenticeship, national traineeship or other governm: 1320
full-time job (over 30 hours a week)                        : 1253
out of work, unemployed                                     :  617
part-time job (if this is your main activity)               :  339
(Other)                                                     :  183
NA's                                                        :   49
                                                         a9_21
pregnancy/looking after children/family                     :   56
waiting to start a new job/government supported training/tra:   40
part-time education                                         :   32
other                                                       :   20
illness/accident                                            :   11
(Other)                                                     :   17
NA's                                                        :14486
```

```
                                                         a9_22
temporary/casual work                                       :    2
waiting to start a new job/government supported training/tra:    2
pregnancy/looking after children/family                     :    1
holiday (school, college, university)                       :    1
not answered (99)                                           :    0
(Other)                                                     :    0
NA's                                                        :14656
                                    a9_23          a11_1a
not answered (99)                      :    0   Length:14662
item not applicable                    :    0   Class :character
part-time education                    :    0   Mode  :character
pregnancy/looking after children/family:    0
temporary/casual work                  :    0
(Other)                                :    0
NA's                                   :14662
   a11_1b             a11_1c             a11_2a             a11_2b
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11_2c             a11_3a             a11_3b             a11_3c
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11_4a             a11_4b             a11_4c             a11_5a
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11_5b             a11_5c             a11_6a             a11_6b
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11_6c             a11_7a             a11_7b             a11_7c
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11_8a             a11_8b             a11_8c             a11_9a
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
   a11_9b             a11_9c            a11_10a            a11_10b
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



  a11_10c            a11_11a            a11_11b            a11_11c
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



  a11_11d                                           a11oga1
Length:14662       sport/physical education studies      :1252
Class :character   drama                                 :1141
Mode  :character   religious studies  (includes theology):1057
                   music                                 : 693
                   information technology/info systems   : 671
                   (Other)                               :6138
                   NA's                                  :3710
                                  a11oga2
religious studies  (includes theology): 750
chemistry                             : 471
sport/physical education studies      : 446
drama                                 : 442
information technology/info systems   : 373
(Other)                               :3792
NA's                                  :8388
                                  a11oga3
physics                               :  435
biology                               :  314
religious studies  (includes theology):  297
chemistry                             :  259
science/single award                  :  200
(Other)                               : 1438
NA's                                  :11719
                                  a11oga4
religious studies  (includes theology):  174
physics                               :  113
chemistry                             :   89
biology                               :   88
science/single award                  :   75
(Other)                               :  662
NA's                                  :13461
                                  a11oga5
religious studies  (includes theology):   80
physics                               :   35
information technology/info systems   :   34
chemistry                             :   30
biology                               :   25
(Other)                               :  233
NA's                                  :14225
```

```
                                   a11oga6
 physics                               :    6
 religious studies  (includes theology):    6
 chemistry                             :    5
 science/single award                  :    3
 statistics                            :    3
 (Other)                               :   17
 NA's                                  :14622
                                   a11oga7
 science in society                    :    2
 biology                               :    1
 mathematics (further)                 :    1
 religious studies  (includes theology):    1
 welsh literature                      :    1
 (Other)                               :    0
 NA's                                  :14656
                                             a11oga8
 stage and performing arts dual award (1st grade):    2
 item not applicable                             :    0
 biology                                         :    0
 biology/human                                   :    0
 biology/ social                                 :    0
 (Other)                                         :    0
 NA's                                            :14660
                     a11oga9                                               a11ogb1

 item not applicable     :    0   sport/physical education studies
:1252
 biology                 :    0   drama                                 :11
 biology/human           :    0   religious studies  (includes theology):10
7
 biology/ social         :    0   music                                 : 6

 biology/human and social:    0   information technology/info systems   : 6
71
 (Other)                 :    0   (Other)                               :61
 NA's                    :14662   NA's                                  :37
                                   a11ogb2
 religious studies  (includes theology): 750
 chemistry                             : 471
 sport/physical education studies      : 446
 drama                                 : 442
 information technology/info systems   : 373
 (Other)                               :3792
 NA's                                  :8388
                                   a11ogb3
 physics                               :  435
 biology                               :  314
 religious studies  (includes theology):  297
 chemistry                             :  259
 science/single award                  :  200
 (Other)                               : 1438
 NA's                                  :11719
                                   a11ogb4
 religious studies  (includes theology):  174
 physics                               :  113
 chemistry                             :   89
 biology                               :   88
 science/single award                  :   75
 (Other)                               :  662
 NA's                                  :13461
```

```
NA's                                  :13461
                                  a11ogb5
religious studies  (includes theology):   80
physics                               :   35
information technology/info systems   :   34
chemistry                             :   30
biology                               :   25
(Other)                               :  233
NA's                                  :14225
                                  a11ogb6
physics                               :    6
religious studies  (includes theology):    6
chemistry                             :    5
science/single award                  :    3
statistics                            :    3
(Other)                               :   17
NA's                                  :14622
                                  a11ogb7
science in society                    :    2
biology                               :    1
mathematics (further)                 :    1
religious studies  (includes theology):    1
welsh literature                      :    1
(Other)                               :    0
NA's                                  :14656
                                            a11ogb8
stage and performing arts dual award (1st grade):    2
item not applicable                             :    0
biology                                         :    0
biology/human                                   :    0
biology/ social                                  :    0
(Other)                                         :    0
NA's                                            :14660
                    a11ogb9        a11ogc1            a11ogc2
item not applicable     :    0   Length:14662       Length:14662
biology                 :    0   Class :character   Class :character
biology/human           :    0   Mode  :character   Mode  :character
biology/ social         :    0
biology/human and social:    0
(Other)                 :    0
NA's                    :14662
  a11ogc3            a11ogc4            a11ogc5            a11ogc6
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




  a11ogc7            a11ogc8            a11ogc9
Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character




                               a11oq1
rsa nvq level 1/certificate       :  214
rsa don't know nvq level/other rsa:  126
gnvq foundation                   :  116
gnvq foundation                   :  110  
gnvq intermediate                 :  104  
other qual, bands unclear         :   77  
(Other)                           :  764  
NA's                              :13261  
                                        a11oq2     
rsa nvq level 1/certificate                :   37  
rsa nvq level 2/diploma                    :   37  
rsa don't know nvq level/other rsa         :   34  
other qual, bands unclear                  :   18  
other band c n.e.c. at nvq level not stated:   13  
(Other)                                    :  118  
NA's                                       :14405  
                                        a11oq3     
rsa don't know nvq level/other rsa         :   13  
rsa nvq level 1/certificate                :   10  
other band c n.e.c. at nvq level not stated:    7  
rsa nvq level 2/diploma                    :    6  
other qual, bands unclear                  :    6  
(Other)                                    :   27  
NA's                                       :14593  
                                        a11oq4     
other band c n.e.c. at nvq level not stated:    2  
rsa nvq level 2/diploma                    :    1  
rsa don't know nvq level/other rsa         :    1  
city & guilds nvq level 1/part 1           :    1  
item not applicable                        :    0  
(Other)                                    :    0  
NA's                                       :14657  
                               a11oq5                           a11oq6     
rsa don't know nvq level/other rsa:    1   item not applicable:    0  
item not applicable               :    0   gcse               :    0  
gcse                              :    0   gcse (short course):    0  
gcse (short course)               :    0   ncc                :    0  
ncc                               :    0   a-level            :    0  
(Other)                           :    0   (Other)            :    0  
NA's                              :14661   NA's               :14662  
                a11oq7                          a11oq8     
item not applicable:    0   item not applicable:    0  
gcse               :    0   gcse               :    0  
gcse (short course):    0   gcse (short course):    0  
ncc                :    0   ncc                :    0  
a-level            :    0   a-level            :    0  
(Other)            :    0   (Other)            :    0  
NA's               :14662   NA's               :14662  
                a11oq9     
item not applicable:    0  
gcse               :    0  
gcse (short course):    0  
ncc                :    0  
a-level            :    0  
(Other)            :    0  
NA's               :14662  
                                           a11os1     
information technology & computer applications:  241  
office and secretarial skills                 :  155  
business                                      :  103  
health & social care                          :  102  
other combined or general courses             :   74  
(Other)                                       :  726  
NA's                                          :13261  
```

```
                                           a11os2
office and secretarial skills                  :   56
information technology & computer applications:   42
mathematics                                    :   13
languages & language studies                   :   12
business & management (general)                :    8
(Other)                                        :  126
NA's                                           :14405
                                           a11os3
information technology & computer applications:   13
office and secretarial skills                  :   11
languages & language studies                   :    6
religious studies                              :    3
travel & tourism                               :    3
(Other)                                        :   33
NA's                                           :14593
                                                         a11os4
communication & mass media                                  :    1
science & technology (general)                              :    1
information technology & computer applications              :    1
not stated                                                  :    1
other - includes child development, hairdressing and beauty :    1
(Other)                                                     :    0
NA's                                                        :14657
                a11os5                           a11os6
not stated         :    1   item not applicable     :    0
item not applicable:    0   biology                 :    0
biology            :    0   biology/human           :    0
biology/human      :    0   biology/social          :    0
biology/social     :    0   biology/human and social:    0
(Other)            :    0   (Other)                 :    0
NA's               :14661   NA's                    :14662
                     a11os7                           a11os8
item not applicable     :    0   item not applicable     :    0
biology                 :    0   biology                 :    0
biology/human           :    0   biology/human           :    0
biology/social          :    0   biology/social          :    0
biology/human and social:    0   biology/human and social:    0
(Other)                 :    0   (Other)                 :    0
NA's                    :14662   NA's                    :14662
                     a11os9         a11or1             a11or2
item not applicable     :    0   Length:14662       Length:14662
biology                 :    0   Class :character   Class :character
biology/human           :    0   Mode  :character   Mode  :character
biology/social          :    0
biology/human and social:    0
(Other)                 :    0
NA's                    :14662
   a11or3             a11or4             a11or5             a11or6
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




   a11or7             a11or8             a11or9
Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

```
                  a12a                                    a12bs1
 not answered (9)   :    0   mathematics                     :  279
 item not applicable:    0   english                         :  211
 yes                : 1109   english literature              :   12
 no                 :11861   science/double award (1st grade):   10
 NA's               : 1692   science/single award            :    5
                             (Other)                         :   66
                             NA's                            :14079
                              a12bs2                         a12bs3
 mathematics                     :   27   science/single award:    9
 english                         :   21   mathematics         :    9
 english literature              :   10   english literature  :    7
 science/single award            :    6   english             :    5
 science/double award (1st grade):    6   geography           :    4
 (Other)                         :   27   (Other)             :   24
 NA's                            :14565   NA's                :14604
                                  a12bs4
 english                             :    6
 art and design                      :    2
 science/single award                :    2
 home economics/food  food technology:    2
 art (without 'design' element)      :    2
 (Other)                             :   20
 NA's                                :14626
                                    a12bs5                       a12bs6
 mathematics                           :    2   science/single award:    1
 business studies                      :    1   mathematics         :    1
 history                               :    1   geography           :    1
 religious studies  (includes theology):    1   other languages     :    1
 english                               :    1   item not applicable :    0
 (Other)                               :    0   (Other)             :    0
 NA's                                  :14656   NA's                :14658
                             a12bs7         a12br1             a12br2

 science/single award          :    1   Length:14662       Length:14662

 cdt/textiles (include textiles):    1   Class :character   Class
:character
 item not applicable           :    0   Mode  :character   Mode  :characte

 biology                       :    0
 biology/human                 :    0
 (Other)                       :    0
 NA's                          :14660
    a12br3             a12br4             a12br5             a12br6
 Length:14662       Length:14662       Length:14662       Length:14662
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character




    a12br7                          a12bx
 Length:14662       not answered (9)   :    0
 Class :character   item not applicable:    0
 Mode  :character   NA's               :14662
```

```
                                                         a12oq1
nvq (not rsa, btec or c & g) level 2                        :  111
nvq (not rsa, btec or c & g) level 1                        :   52
gnvq intermediate                                           :   46
other qualification: band not known, i.e. all other courses :   32
gce a-level                                                 :   29
(Other)                                                     :  274
NA's                                                        :14118
                                                         a12oq2
nvq (not rsa, btec or c & g) level 2                        :   21
rsa nvq level 1/certificate                                 :   10
other qualification: band not known, i.e. all other courses :   10
gce a-level                                                 :    9
nvq (not rsa, btec or c & g) level 3                        :    9
(Other)                                                     :   52
NA's                                                        :14551
                                 a12oq3
nvq (not rsa, btec or c & g) level 1:    5
rsa nvq level 1/certificate         :    4
gce a-level                         :    2
nvq (not rsa, btec or c & g) level 3:    2
gnvq intermediate                   :    1
(Other)                             :    7
NA's                                :14641
                                 a12oq4
gce a-level                         :    1
nvq (not rsa, btec or c & g) level 2:    1
other band c n.e.c. at nvq level 3  :    1
item not applicable                 :    0
gcse                                :    0
(Other)                             :    0
NA's                                :14659
                                  a12oq5
other band c n.e.c. at nvq level 3   :    1
item not applicable                  :    0
gcse                                 :    0
gcse short course (specific mentions):    0
ncc (national curriculum certificate):    0
(Other)                              :    0
NA's                                 :14661
                                  a12oq6
other band c n.e.c. at nvq level 3   :    1
item not applicable                  :    0
gcse                                 :    0
gcse short course (specific mentions):    0
ncc (national curriculum certificate):    0
(Other)                              :    0
NA's                                 :14661
                                           a12os1
information technology & computer applications:  108
business & management (general)               :   45
business                                      :   33
office and secretarial skills                 :   31
health & social care                          :   22
(Other)                                       :  330
NA's                                          :14093
```

```
                                           a12os2
information technology & computer applications:   27
office and secretarial skills                 :   13
business & management (general)               :    9
hotel & commercial catering                   :    4
mathematics                                   :    3
(Other)                                       :   56
NA's                                          :14550
                                           a12os3
information technology & computer applications:    3
office and secretarial skills                 :    2
hotel & commercial catering                   :    2
biology/human                                 :    1
chemistry                                     :    1
(Other)                                       :   12
NA's                                          :14641
                                     a12os4                      a12os5
religious studies (includes theology)   :    1   not stated         :    1

health (general) & health administration:    1   not answered (999) :    0

not stated                              :    1   item not applicable:    0

not answered (999)                      :    0   biology            :    0
item not applicable                     :    0   biology/human      :    0

(Other)                                 :    0   (Other)            :    0
NA's                                    :14659   NA's               :14661
                a12os6         a12or1             a12or2
not stated         :    1   Length:14662       Length:14662
not answered (999) :    0   Class :character   Class :character
item not applicable:    0   Mode  :character   Mode  :character
biology            :    0
biology/human      :    0
(Other)            :    0
NA's               :14661
   a12or3             a12or4             a12or5             a12or6
Length:14662       Length:14662       Length:14662       Length:14662
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




                a12ox                        a13a
not answered (9)   :    0   not answered (9)    :    0
item not applicable:    0   item not applicable:    0
NA's               :14662   yes                : 1911
                            no                 :11210
                            NA's               : 1541



                                 a13bq1
gce a-level                         :  819
gcse                                :  280
nvq (not rsa, btec or c & g) level 2:  101
gnvq intermediate                   :   97
other gnvq (not codes 08-12)        :   87
(Other)                             :  480
NA's                                :12798
```

```
                                 a13bq2
gce a-level                         :  304
gcse                                :   66
nvq (not rsa, btec or c & g) level 2:   15
nvq (not rsa, btec or c & g) level 3:    8
gce a/s exam                        :    7
(Other)                             :   41
NA's                                :14221
                         a13bq3
gce a-level                 :  195
gcse                        :   40
other gnvq (not codes 08-12):    3
gce a/s exam                :    2
gnvq advanced               :    2
(Other)                     :   13
NA's                        :14407
                                           a13bq4
gcse                                          :   17
gce a-level                                   :   14
gce a/s exam                                  :    2
rsa nvq level 1/certificate                   :    1
city & guilds don't know nvq level/other c & g:    1
(Other)                                       :    1
NA's                                          :14626
                                  a13bq5
schedule not obtained                :    0
schedule not applicable              :    0
item not applicable                  :    0
gcse                                 :    0
gcse short course (specific mentions):    0
(Other)                              :    0
NA's                                 :14662
                                  a13bq6
gcse                                 :    1
item not applicable                  :    0
gcse short course (specific mentions):    0
ncc (national curriculum certificate):    0
gce a-level                          :    0
(Other)                              :    0
NA's                                 :14661
                                                         a13bs1
mathematics                                                 :  171
business                                                    :   81
english                                                     :   77
biology                                                     :   68
other includes child development, hairdressing and beauty ca:   67
(Other)                                                     : 1447
NA's                                                        :12751
               a13bs2                   a13bs3                  a13bs4
mathematics       :   33   mathematics     :   22   mathematics    :    4
english           :   29   sociology       :   20   general studies:    4
biology           :   22   business studies:   13   history        :    3
geography         :   22   geography       :   13   law            :    2
english literature:   22   english         :   13   psychology     :    2
(Other)           :  324   (Other)         :  176   (Other)        :   22
NA's              :14210   NA's            :14405   NA's           :14625
               a13bs5                                        a13bs6
physics           :    1   science: double award (1st grade):    1
business studies  :    1   item not applicable              :    0
art and design    :    1   biology                          :    0
```

```
 sociology          :    1   biology: human                  :    0
 english literature:    1   biology: social                 :    0
 (Other)            :    4   (Other)                         :    0
 NA's               :14653   NA's                            :14661
                 a13c
 not answered (9)   :    0
 item not applicable:    0
 yes                :  944
 no                 :  945
 NA's               :12773


                                         a13d
 college of further education (state system):  390
 sixth form college (state system)          :  219
 state school (including grant maintained)  :  141
 training centre                            :  102
 independent/other college                  :   49
 (Other)                                    :   13
 NA's                                       :13748
                a14_1                        a14_2
 not answered (9)   :    0   not answered (9)   :    0
 item not applicable:    0   item not applicable:    0
 yes                :11053   yes                :10773
 no                 : 3465   no                 :  205
 NA's               :  144   NA's               : 3684


                a15_1              a15_2          a15_21          a15_22
 not answered (9)   :   0   Min.   :1      Min.   :0.000   Min.   :0.000
 item not applicable:   0   1st Qu.:1      1st Qu.:0.000   1st Qu.:0.000
 yes                :5388   Median :1      Median :0.000   Median :0.000
 no                 :5618   Mean   :1      Mean   :0.045   Mean   :0.021
 NA's               :3656   3rd Qu.:1      3rd Qu.:0.000   3rd Qu.:0.000
                            Max.   :1      Max.   :1.000   Max.   :1.000
                            NA's   :9289   NA's   :9289    NA's   :9289
     a15_23          a15_24          a15_25         a15_26           a16

 Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   : 1.0
0
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:16.
00
 Median :0.000   Median :1.000   Median :0.00   Median :0.000   Median :20.
00
 Mean   :0.301   Mean   :0.595   Mean   :0.02   Mean   :0.018   Mean   :21.0
9
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.00   3rd Qu.:0.000   3rd Qu.:25.
00
 Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.000   Max.   :50.0
0
 NA's   :9289    NA's   :9289    NA's   :9289   NA's   :9289    NA's   :416

                 a17                         a18_1          a18_2a
 not answered (9)   :    0   not answered (9)   :   0   Min.   :   5.0
 item not applicable:    0   item not applicable:   0   1st Qu.: 130.0
 yes                :4848   yes                : 803   Median : 180.0
 no                 :4137   no                 :9999   Mean   : 286.5
 not sure           :1976   NA's               :3860   3rd Qu.: 400.0
 NA's               :3701                              Max.   :3000.0
                                                       NA's   :13939

                a18_2b              a19_1              a20_1
                 a18_2b             a19_1            a20_1
 not answered (9)   :    0   sep       :9948   jun       :5042
 item not applicable:    0   aug       : 363   jul       :3189
 term               :  435   dk/na date: 267   may       :1029
 year               :  298   oct       : 147   don't know: 577
 other period       :   11   jul       :  73   2000      : 282
 NA's               :13918   (Other)   : 255   (Other)   : 934
                             NA's      :3609   NA's      :3609
                 a21_1
 not answered (9)   :    0
 item not applicable:    0
 yes                :  868
 no                 :12737
 NA's               : 1057



                                         a21_2a
 college of further education (state system):  576
 somewhere else                             :   86
 work                                       :   52
 training centre run by your employer       :   48
 private training centre                    :   41
 (Other)                                    :   56
 NA's                                       :13803
                                                     a21_2b
 there are no course fees to pay                        :  232
 my parents/family/me                                   :  208
 it is paid for some other way                          :  166
 my employer                                            :  163
 it is paid for using a training voucher or plastic card:   55
 (Other)                                                :    0
 NA's                                                   :13838
                  a22                       a22gcse
 not answered (9)   :   0   not answered (9)   :   0
 item not applicable:   0   item not applicable:   0
 yes                :8008   yes                :2026
 no                 :6282   no                 :5243
 NA's               : 372   NA's               :7393



                              a22a1                       a22a2
 mathematics                     :  755   mathematics       :  176
 english                         :  516   english           :  153
 spanish                         :   72   english literature:   33
 general studies                 :   60   sociology         :   24
 sport/physical education studies:   41   biology           :   13
 (Other)                         :  582   (Other)           :  207
 NA's                            :12636   NA's              :14056
                  a22a3                                            a22a4
 mathematics         :   54   mathematics                        :   15
 science/single award:   24   information technology/info systems:   14
 english             :   18   sociology                          :   14
 biology             :   16   english                            :   14
 sociology           :   11   biology                            :   12
 (Other)             :  157   (Other)                            :  113
 NA's                :14382   NA's                               :14480
              a22a5                       a22a6
 mathematics     :   10   mathematics         :    4
 english         :   10   geography           :    3
 biology         :    5   english             :    3
 sociology       :    5   chemistry           :    2
 sociology       :    5   chemistry            :    2
 business studies:    4   science/single award:    2
 (Other)         :   57   (Other)             :   19
 NA's            :14571   NA's                :14629
                 a22as                                               a22b1
 not answered (9)   :   0   general studies                         :  144
 item not applicable:   0   mathematics                             :   73
 yes                : 788   mathematics (further)                   :   41
 no                 :6230   religious studies  (includes theology):   40
 NA's               :7644   french                                  :   37
                            (Other)                                 :  453
                            NA's                                    :13874
                                    a22b2
 general studies                        :   14
 statistics                             :   10
 religious studies  (includes theology):    7
 computer studies/computing             :    6
 sociology                              :    6
 (Other)                                :   64
 NA's                                   :14555
                               a22b3
 mathematics (pure and statistics):    2
 music                            :    2
 general studies                  :    2
 biology                          :    1
 science/single award             :    1
 (Other)                          :    8
 NA's                             :14646
                                 a22b4                      a22b5
 information technology/info systems:    1   history             :    1
 geography                          :    1   not answered (999) :    0
 not answered (999)                 :    0   item not applicable:    0
 item not applicable                :    0   biology             :    0
 biology                            :    0   biology/human       :    0
 (Other)                            :    0   (Other)             :    0
 NA's                               :14660   NA's                :14661
                 a22b6                     a22alev
 not answered (999) :    0   not answered (9)   :   0
 item not applicable:    0   item not applicable:   0
 biology            :    0   yes                :6939
 biology/human      :    0   no                 : 882
 biology/social     :    0   NA's               :6841
 (Other)            :    0
 NA's               :14662
                a22c1                     a22c2                     a22c3
 mathematics       : 874   chemistry         : 577   geography         : 42

 biology           : 692   biology           : 521   mathematics       : 40

 english literature: 577   physics           : 448   biology           : 34

 geography         : 444   geography         : 426   english literature: 33

 chemistry         : 370   english literature: 414   physics           : 29

 (Other)           :3982   (Other)           :4193   (Other)           :415

 NA's              :7723   NA's              :8083   NA's              :872


                   a22c4                                  a22c5
 general studies       :  576   general studies              :   47
 general studies      :  570   general studies                :   17
 chemistry            :   57   mathematics (further)          :    4
 mathematics (further):   50   art (without 'design' element):    2
 physics              :   36   biology                        :    1
 biology              :   30   physics                        :    1
 (Other)              :  436   (Other)                        :   11
 NA's                 :13477   NA's                           :14596
                                 a22c6                         a23
 information technology/info systems:    1   not answered (9)   :    0
 technology                         :    1   item not applicable:    0
 not answered (999)                 :    0   yes                : 2430
 item not applicable                :    0   no                 :11675
 biology                            :    0   NA's               :  557
 (Other)                            :    0
 NA's                               :14660
    a231a                          a231b
 Length:14662       not answered (9)   :    0
 Class :character   item not applicable:    0
 Mode  :character   full award         :  182
                    certain units only :   38
                    NA's               :14442



                         a231c1                               a231c2
 health & social care       :   58   information technology (it):    2
 business                   :   48   leisure and tourism        :    2
 not stated                 :   27   science                    :    2
 engineering                :   20   business                   :    1
 information technology (it):   19   health & social care       :    1
 (Other)                    :   61   (Other)                    :    2
 NA's                       :14429   NA's                       :14652
                 a231c3                       a231c4
 leisure and tourism:    2   not answered (999) :    0
 other gnvq         :    2   item not applicable:    0
 not answered (999) :    0   performing arts    :    0
 item not applicable:    0   art and design     :    0
 performing arts    :    0   business           :    0
 (Other)            :    0   (Other)            :    0
 NA's               :14658   NA's               :14662
                 a231c5                       a231c6         a232a
 not answered (999) :    0   not answered (999) :    0   Length:14662
 item not applicable:    0   item not applicable:    0   Class :character
 performing arts    :    0   performing arts    :    0   Mode  :character
 art and design     :    0   art and design     :    0
 business           :    0   business           :    0
 (Other)            :    0   (Other)            :    0
 NA's               :14662   NA's               :14662
                 a232b                         a232c1
 not answered (9)   :    0   business             :  298
 item not applicable:    0   health & social care:  254
 full award         : 1073   leisure and tourism :  209
 certain units only :  119   not stated           :  112
 NA's               :13470   art and design       :  108
                             (Other)              :  242
                             NA's                 :13439
                              a232c2                          a232c3
 information technology (it)     :    3   other gnvq         :    1
 leisure and tourism             :    2   not answered (999) :    0
 other gnvq                      :    2   item not applicable:    0
 science                         :    1   performing arts    :    0
 retail and distributive services:    1   art and design     :    0
```

```
 (Other)                          :    0   (Other)             :    0  
 NA's                             :14653   NA's                :14661  
                a232c4                       a232c5     
 not answered (999) :    0   not answered (999) :    0  
 item not applicable:    0   item not applicable:    0  
 performing arts    :    0   performing arts    :    0  
 art and design     :    0   art and design     :    0  
 business           :    0   business           :    0  
 (Other)            :    0   (Other)            :    0  
 NA's               :14662   NA's               :14662  
                a232c6          a233a                              a233b      
 not answered (999) :    0   Length:14662       not answered (9)   :    0  
 item not applicable:    0   Class :character   item not applicable:    0  
 performing arts    :    0   Mode  :character   full award         :  940  
 art and design     :    0                      certain units only :   26  
 business           :    0                      NA's               :13696  
 (Other)            :    0                                                 
 NA's               :14662                                                 
                        a233c1                                    a233c2     
 business                    :  301   leisure and tourism             :    2  
                                                                              
 leisure and tourism         :  166   retail and distributive services:    1  
                                                                              
 health & social care        :  163   not answered (999)              :    0  
                                                                              
 art and design              :   88   item not applicable             :    0  
                                                                              
 information technology (it):   64   performing arts                 :    0  
                                                                              
 (Other)                     :  197   (Other)                         :    0  
 NA's                        :13683   NA's                            :14659  
                        a233c3                       a233c4     
 information technology (it):    1   other gnvq         :    1  
 not answered (999)          :    0   not answered (999) :    0  
 item not applicable         :    0   item not applicable:    0  
 performing arts             :    0   performing arts    :    0  
 art and design              :    0   art and design     :    0  
 (Other)                     :    0   (Other)            :    0  
 NA's                        :14661   NA's               :14661  
                a233c5                       a233c6     
 other gnvq         :    1   other gnvq         :    1  
 not answered (999) :    0   not answered (999) :    0  
 item not applicable:    0   item not applicable:    0  
 performing arts    :    0   performing arts    :    0  
 art and design     :    0   art and design     :    0  
 (Other)            :    0   (Other)            :    0  
 NA's               :14661   NA's               :14661  
                a24_1          a24nva1      
 not answered (9)   :    0   Min.   :0.0000  
 item not applicable:    0   1st Qu.:0.0000  
 yes                : 3243   Median :0.0000  
 no                 :10696   Mean   :0.1506  
 NA's               :  723   3rd Qu.:0.0000  
                             Max.   :3.0000  
                                                                          
                                                                 a24n11     
 other - includes child development, hairdressing and beauty :  292  
 business & management (general)                             :  272  
 hotel & commercial catering                                 :  124  
 marketing sales & distribution                              :  123  
 vehicle maintenance & repair                                      :  122
 (Other)                                                           :  856
 NA's                                                              :12873
                            a24n12
 languages & language studies   :    2
 business & management (general):    1
 office and secretarial skills  :    1
 design (non-industrial)        :    1
 religious studies              :    1
 (Other)                        :    4
 NA's                           :14652
                                   a24n13
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24n14
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24n15
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24n16
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24n17
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24n18          a24nvl1          a24nvl11

 item not applicable                   :    0   Min.   :1.000   Min.   :0.0
0
 business & management (general)       :    0   1st Qu.:1.000   1st
Qu.:0.000
 enterprises                           :    0   Median :1.000   Median
:0.000
 management skills systems & techniques:    0   Mean   :1.234   Mean
```

```
:0.211
 human resources management            :    0   3rd Qu.:1.000   3rd Qu.:0.0
0
 (Other)                               :    0   Max.   :3.000   Max.   :1.0

 NA's                                  :14662   NA's   :12873   NA's   :128

    a24nvl12        a24nvl13        a24nvl18        a24nvl19    
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
 Median :1.000   Median :0.000   Median :0.000   Median :0.000  
 Mean   :0.654   Mean   :0.228   Mean   :0.108   Mean   :0.033  
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000  
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000  
 NA's   :12873   NA's   :12873   NA's   :12873   NA's   :12873  
    a24nva2        
 Min.   :0.000000  
 1st Qu.:0.000000  
 Median :0.000000  
 Mean   :0.009276  
 3rd Qu.:0.000000  
 Max.   :3.000000  


                                                           a24n21     
 business & management (general)                             :   17  
 other - includes child development, hairdressing and beauty :   15  
 office and secretarial skills                                :   12  
 marketing sales & distribution                               :    8  
 hotel & commercial catering                                  :    8  
 (Other)                                                      :   53  
 NA's                                                         :14549  
                                   a24n22     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management            :    0  
 (Other)                               :    0  
 NA's                                  :14662  
                                   a24n23     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management            :    0  
 (Other)                               :    0  
 NA's                                  :14662  
                                   a24n24     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management            :    0  
 (Other)                               :    0  
 NA's                                  :14662  
                                   a24n25     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24n26
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24n27
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24n28            a24nvl2          a24nvl21

 item not applicable                    :    0   Min.   :1.000   Min.   :0.0

 business & management (general)        :    0   1st Qu.:1.000   1st
Qu.:0.00
 enterprises                            :    0   Median :1.000   Median :0.0
 management skills systems & techniques:    0   Mean   :1.088   Mean
:0.15
 human resources management             :    0   3rd Qu.:1.000   3rd Qu.:0.0

 (Other)                                :    0   Max.   :3.000   Max.   :1.0

 NA's                                   :14662   NA's   :14549   NA's   :145

    a24nvl22        a24nvl23        a24nvl28        a24nvl29
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
 Median :1.000   Median :0.000   Median :0.000   Median :0.000
 Mean   :0.513   Mean   :0.363   Mean   :0.053   Mean   :0.009
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
 NA's   :14549   NA's   :14549   NA's   :14549   NA's   :14549
    a24nva3                                                        a24n31
 Min.   :0.000000   information technology & computer applications:    5
 1st Qu.:0.000000   office and secretarial skills                  :    2
 Median :0.000000   financial management & accounting              :    1
 Mean   :0.001432   public administration                          :    1
 3rd Qu.:0.000000   languages & language studies                   :    1
 Max.   :3.000000   (Other)                                        :    6
                    NA's                                           :14646
                                    a24n32
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24n33
 item not applicable                    :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24n34
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24n35
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24n36
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24n37
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24n38        a24nvl3          a24nvl31

 item not applicable                   :    0   Min.   :1.000   Min.   :0.0

 business & management (general)       :    0   1st Qu.:1.000   1st
Qu.:0.00
 enterprises                           :    0   Median :1.000   Median :0.0
 management skills systems & techniques:    0   Mean   :1.125   Mean
:0.25
 human resources management            :    0   3rd Qu.:1.000   3rd Qu.:0.2

 (Other)                               :    0   Max.   :3.000   Max.   :1.0

 NA's                                  :14662   NA's   :14646   NA's   :146

    a24nvl32        a24nvl33        a24nvl38        a24nvl39
 Min.   :0.000   Min.   :0.00    Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.00    1st Qu.:0.000   1st Qu.:0.000
 Median :0.000   Median :0.00    Median :0.000   Median :0.000
 Mean   :0.312   Mean   :0.25    Mean   :0.188   Mean   :0.125
 3rd Qu.:1.000   3rd Qu.:0.25    3rd Qu.:0.000   3rd Qu.:0.000
 Max.   :1.000   Max.   :1.00    Max.   :1.000   Max.   :1.000
 NA's   :14646   NA's   :14646   NA's   :14646   NA's   :14646
```

```
NA's   :14646   NA's   :14646   NA's   :14646   NA's   :14646  
  a24bta1         
 Length:14662      
 Class :character  
 Mode  :character  
                   
                   
                   
                   
                                                         a24b11     
 engineering/ technology/ manufacture (general)             :   20  
 mechanical engineering                                     :    6  
 vehicle maintenance & repair                               :    6  
 other - includes child development, hairdressing and beauty :    6  
 public administration                                      :    5  
 (Other)                                                    :   31  
 NA's                                                       :14588  
                                   a24b12     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 management skills systems & techniques:    0  
 human resources management               :    0  
 (Other)                                  :    0  
 NA's                                     :14662  
                                   a24b13     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 management skills systems & techniques:    0  
 human resources management               :    0  
 (Other)                                  :    0  
 NA's                                     :14662  
                                   a24b14     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 management skills systems & techniques:    0  
 human resources management               :    0  
 (Other)                                  :    0  
 NA's                                     :14662  
                                   a24b15     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 management skills systems & techniques:    0  
 human resources management               :    0  
 (Other)                                  :    0  
 NA's                                     :14662  
                                   a24b16     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 management skills systems & techniques:    0  
 human resources management               :    0  
 (Other)                                  :    0  
 NA's                                     :14662  
                                   a24b17     
 item not applicable                      :    0  
 business & management (general)          :    0  
 enterprises                              :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24b18         a24btc1         a24btc11    

 item not applicable                  :    0   Min.   :1       Min.   
:0.000  
 business & management (general)      :    0   1st Qu.:1       1st 
Qu.:0.000  
 enterprises                          :    0   Median :1       Median 
:0.000  
 management skills systems & techniques:    0   Mean   :1       Mean   :0.10
1  
 human resources management           :    0   3rd Qu.:1       3rd Qu.:0.00
0  
 (Other)                              :    0   Max.   :1       Max.   :1.00

 NA's                                 :14662   NA's   :14593   NA's   :1459

    a24btc12        a24btc13        a24btc18        a24btc19    
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0      
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0      
 Median :0.000   Median :0.000   Median :1.000   Median :0      
 Mean   :0.116   Mean   :0.029   Mean   :0.754   Mean   :0      
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0      
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0      
 NA's   :14593   NA's   :14593   NA's   :14593   NA's   :14593  
   a24bta2                                                       a24b21     
 Length:14662       veterinary services & pet care               :   18  
 Class :character   public administration                        :   13  
 Mode  :character   theatre & dramatic arts                      :   10  
                    agricultural & horticultural studies (general):    7  
                    information technology & computer applications:    7  
                    (Other)                                      :   58  
                    NA's                                         :14549  
                                   a24b22      
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24b23      
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24b24      
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
 NA's                                  :14662
                                    a24b25
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24b26
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24b27
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24b28         a24btc2         a24btc21

 item not applicable                   :    0   Min.   :1.00    Min.   
:0.000
 business & management (general)       :    0   1st Qu.:1.00    1st 
Qu.:0.000
 enterprises                           :    0   Median :1.00    Median 
:0.000
 management skills systems & techniques:    0   Mean   :1.01    Mean   
:0.072
 human resources management            :    0   3rd Qu.:1.00    3rd Qu.:0.0
0
 (Other)                               :    0   Max.   :2.00    Max.   :1.0

 NA's                                  :14662   NA's   :14565   NA's   :145

    a24btc22        a24btc23        a24btc28        a24btc29
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0  
 Median :0.000   Median :0.000   Median :1.000   Median :0  
 Mean   :0.072   Mean   :0.041   Mean   :0.825   Mean   :0  
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0  
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0  
 NA's   :14565   NA's   :14565   NA's   :14565   NA's   :14565  
   a24bta3
 Length:14662
 Class :character
 Mode  :character




                                                          a24b31
 other - includes child development, hairdressing and beauty :   73
 theatre & dramatic arts                                     :   54
 sports studies & combined sports                            :   47
 sports studies & combined sports                          :   47
 design (non-industrial)                                   :   34
 information technology & computer applications            :   34
 (Other)                                                   :  343
 NA's                                                      :14077
                          a24b32
 languages & language studies:    1
 theatre & dramatic arts     :    1
 mathematics                 :    1
 electronic engineering      :    1
 item not applicable         :    0
 (Other)                     :    0
 NA's                        :14658
                                           a24b33
 information technology & computer applications:    1
 item not applicable                           :    0
 business & management (general)               :    0
 enterprises                                   :    0
 management skills systems & techniques        :    0
 (Other)                                       :    0
 NA's                                          :14661
                                   a24b34
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24b35
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24b36
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24b37
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                   a24b38         a24btc3         a24btc31

 item not applicable                   :    0   Min.   :1.000   Min.   :0.00
0
 business & management (general)       :    0   1st Qu.:1.000   1st
Qu.:0.000
 enterprises                           :    0   Median :1.000   Median
:0.000
```

```
 management skills systems & techniques:    0   Mean   :1.009   Mean   
:0.013  
 human resources management            :    0   3rd Qu.:1.000   3rd Qu.:0.0
0  
 (Other)                               :    0   Max.   :3.000   Max.   :1.0
                                                                          
 NA's                                  :14662   NA's   :14123   NA's   :141
                                                                          
    a24btc32        a24btc33        a24btc38        a24btc39    
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0      
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0      
 Median :0.000   Median :0.000   Median :1.000   Median :0      
 Mean   :0.048   Mean   :0.096   Mean   :0.852   Mean   :0      
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0      
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0      
 NA's   :14123   NA's   :14123   NA's   :14123   NA's   :14123  
   a24bta4         
 Length:14662      
 Class :character  
 Mode  :character  
                   
                   
                   
                   
                                                          a24b41     
 other - includes child development, hairdressing and beauty :    4  
 design (non-industrial)                                     :    3  
 mathematics                                                 :    3  
 mechanical engineering                                      :    3  
 not stated                                                  :    3  
 (Other)                                                     :   22  
 NA's                                                        :14624  
                                    a24b42     
 earth sciences                         :    1  
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 (Other)                                :    0  
 NA's                                   :14661  
                                    a24b43     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24b44     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24b45     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
```

```
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24b46
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24b47
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24b48        a24btc4         a24btc41

item not applicable                   :    0   Min.   :1.000   Min.   :0

business & management (general)       :    0   1st Qu.:1.000   1st Qu.:0

enterprises                           :    0   Median :1.000   Median :0

management skills systems & techniques:    0   Mean   :1.034   Mean   :0

human resources management            :    0   3rd Qu.:1.000   3rd Qu.:0

(Other)                               :    0   Max.   :2.000   Max.   :0

NA's                                  :14662   NA's   :14633   NA's   :146

   a24btc42        a24btc43        a24btc48        a24btc49
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0
Median :0.000   Median :0.000   Median :1.000   Median :0
Mean   :0.069   Mean   :0.069   Mean   :0.897   Mean   :0
3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0
Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0
NA's   :14633   NA's   :14633   NA's   :14633   NA's   :14633
  a24cga1
Length:14662
Class :character
Mode  :character




                                                        a24c11
electrical engineering                                    :   22
vehicle maintenance & repair                              :   18
electronic engineering                                    :   16
engineering/ technology/ manufacture (general)            :   13
other - includes child development, hairdressing and beauty :   10
(Other)                                                   :   68
NA's                                                      :14515
```

```
                                        a24c12
 languages & language studies         :    1
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 (Other)                              :    0
 NA's                                 :14661
                                        a24c13
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                        a24c14
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                        a24c15
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                        a24c16
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                        a24c17
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                        a24c18         a24cgc1         a24cgc11

 item not applicable                  :    0   Min.   :1.000   Min.   :0.0
0
 business & management (general)      :    0   1st Qu.:1.000   1st
Qu.:0.000
 enterprises                          :    0   Median :1.000   Median
:0.000
 management skills systems & techniques:    0   Mean   :1.101   Mean
:0.295
 human resources management           :    0   3rd Qu.:1.000   3rd Qu.:1.0
0
```

```
(Other)                               :    0   Max.   :3.000   Max.   :1.0

NA's                                  :14662   NA's   :14533   NA's   :145

   a24cgc12        a24cgc13        a24cgc18        a24cgc19    
 Min.   :0.000   Min.   :0.00    Min.   :0.000   Min.   :0     
 1st Qu.:0.000   1st Qu.:0.00    1st Qu.:0.000   1st Qu.:0     
 Median :0.000   Median :0.00    Median :1.000   Median :0     
 Mean   :0.147   Mean   :0.07    Mean   :0.589   Mean   :0     
 3rd Qu.:0.000   3rd Qu.:0.00    3rd Qu.:1.000   3rd Qu.:0     
 Max.   :1.000   Max.   :1.00    Max.   :1.000   Max.   :0     
 NA's   :14533   NA's   :14533   NA's   :14533   NA's   :14533  
  a24cga2                                                     a24c21     
 Length:14662       electrical engineering                        :   10  
 Class :character   engineering/ technology/ manufacture (general):   10  
 Mode  :character   mathematics                                   :    5  
                    vehicle maintenance & repair                  :    5  
                    not stated                                    :    5  
                    (Other)                                       :   33  
                    NA's                                          :14594  
                                   a24c22     
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24c23     
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24c24     
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24c25     
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
                                   a24c26     
 item not applicable                  :    0  
 business & management (general)      :    0  
 enterprises                          :    0  
 management skills systems & techniques:    0  
 human resources management           :    0  
 (Other)                              :    0  
 NA's                                 :14662  
```

```
                                        a24c27
 item not applicable                      :    0
 business & management (general)          :    0
 enterprises                              :    0
 management skills systems & techniques:    0
 human resources management               :    0
 (Other)                                  :    0
 NA's                                     :14662
                                        a24c28        a24cgc2         a24cgc21

 item not applicable                      :    0   Min.   :1.000   Min.   :0.0(
0
 business & management (general)          :    0   1st Qu.:1.000   1st
Qu.:0.000
 enterprises                              :    0   Median :1.000   Median
:0.000
 management skills systems & techniques:    0   Mean   :1.033   Mean
:0.033
 human resources management               :    0   3rd Qu.:1.000   3rd Qu.:0.0(
0
 (Other)                                  :    0   Max.   :2.000   Max.   :1.0(

 NA's                                     :14662   NA's   :14602   NA's   :146(

    a24cgc22        a24cgc23        a24cgc28        a24cgc29
 Min.   :0.000   Min.   :0.000   Min.   :0.00    Min.   :0
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00    1st Qu.:0
 Median :0.000   Median :0.000   Median :1.00    Median :0
 Mean   :0.267   Mean   :0.083   Mean   :0.65    Mean   :0
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:1.00    3rd Qu.:0
 Max.   :1.000   Max.   :1.000   Max.   :1.00    Max.   :0
 NA's   :14602   NA's   :14602   NA's   :14602   NA's   :14602
   a24cga3                                                  a24c31
 Length:14662       vehicle maintenance & repair          :    3
 Class :character   not stated                            :    3
 Mode  :character   cooking & food & drinking preparation :    2
                    mathematics                           :    2
                    building/construction studies, general:    2
                    (Other)                               :    8
                    NA's                                  :14642
                                        a24c32
 item not applicable                      :    0
 business & management (general)          :    0
 enterprises                              :    0
 management skills systems & techniques:    0
 human resources management               :    0
 (Other)                                  :    0
 NA's                                     :14662
                                        a24c33
 item not applicable                      :    0
 business & management (general)          :    0
 enterprises                              :    0
 management skills systems & techniques:    0
 human resources management               :    0
 (Other)                                  :    0
 NA's                                     :14662
                                        a24c34
 item not applicable                      :    0
 business & management (general)          :    0
 enterprises                              :    0
 management skills systems & techniques:    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24c35
 item not applicable                    :    0
 business & management (general)       :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24c36
 item not applicable                    :    0
 business & management (general)       :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24c37
 item not applicable                    :    0
 business & management (general)       :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                    a24c38         a24cgc3         a24cgc31

 item not applicable                    :    0   Min.   :1       Min.   
:0.000
 business & management (general)       :    0   1st Qu.:1       1st 
Qu.:0.000
 enterprises                            :    0   Median :1       Median 
:0.000
 management skills systems & techniques:    0   Mean   :1       Mean   :0.0
6
 human resources management             :    0   3rd Qu.:1       3rd Qu.:0.0
0
 (Other)                                :    0   Max.   :1       Max.   :1.0

 NA's                                   :14662   NA's   :14644   NA's   :146

    a24cgc32         a24cgc33         a24cgc38         a24cgc39     
 Min.   :0       Min.   :0.000   Min.   :0.000   Min.   :0  
 1st Qu.:0       1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0  
 Median :0       Median :0.000   Median :1.000   Median :0  
 Mean   :0       Mean   :0.389   Mean   :0.556   Mean   :0  
 3rd Qu.:0       3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0  
 Max.   :0       Max.   :1.000   Max.   :1.000   Max.   :0  
 NA's   :14644   NA's   :14644   NA's   :14644   NA's   :14644  
   a24cga4         
 Length:14662      
 Class :character  
 Mode  :character  



                                                         a24c41
```

```
                                                          a24c41
 other - includes child development, hairdressing and beauty :   32
 information technology & computer applications              :   21
 mathematics                                                 :   14
 vehicle maintenance & repair                                :   14
 not stated                                                  :   13
 (Other)                                                     :   87
 NA's                                                        :14481
                             a24c42
 languages & language studies   :    2
 mathematics                    :    1
 item not applicable            :    0
 business & management (general):    0
 enterprises                    :    0
 (Other)                        :    0
 NA's                           :14659
                                    a24c43
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24c44
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24c45
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24c46
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24c47
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24c48         a24cgc4         a24cgc41

 item not applicable                   :    0   Min.   :1.00    Min.
:0.000
 business & management (general)       :    0   1st Qu.:1.00    1st
 business & management (general)        :    0   1st Qu.:1.00    1st
Qu.:0.000
 enterprises                            :    0   Median :1.00    Median
:0.000
 management skills systems & techniques:    0   Mean   :1.14    Mean
:0.134
 human resources management             :    0   3rd Qu.:1.00    3rd Qu.:0.0
0
 (Other)                                :    0   Max.   :3.00    Max.   :1.0

 NA's                                   :14662   NA's   :14498   NA's   :144

    a24cgc42         a24cgc43         a24cgc48         a24cgc49
 Min.   :0.000   Min.   :0.00    Min.   :0.000   Min.   :0
 1st Qu.:0.000   1st Qu.:0.00    1st Qu.:0.000   1st Qu.:0
 Median :0.000   Median :0.00    Median :1.000   Median :0
 Mean   :0.183   Mean   :0.14    Mean   :0.683   Mean   :0
 3rd Qu.:0.000   3rd Qu.:0.00    3rd Qu.:1.000   3rd Qu.:0
 Max.   :1.000   Max.   :1.00    Max.   :1.000   Max.   :0
 NA's   :14498   NA's   :14498   NA's   :14498   NA's   :14498
   a24rsa1                                                        a24r11
 Length:14662       information technology & computer applications:  142
 Class :character   office and secretarial skills                 :   94
 Mode  :character   not stated                                    :   15
                    business & management (general)               :   14
                    languages & language studies                  :    7
                    (Other)                                       :   13
                    NA's                                          :14377
                                                   a24r12
 information technology & computer applications:    3
 languages & language studies                  :    2
 office and secretarial skills                 :    1
 item not applicable                           :    0
 business & management (general)               :    0
 (Other)                                       :    0
 NA's                                          :14656
                                           a24r13
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                           a24r14
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                           a24r15
 item not applicable                    :    0
 business & management (general)        :    0
 enterprises                            :    0
 management skills systems & techniques:    0
 human resources management             :    0
 (Other)                                :    0
 NA's                                   :14662
                                           a24r16
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management            :    0  
 (Other)                               :    0  
 NA's                                  :14662  
                                    a24r17     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management            :    0  
 (Other)                               :    0  
 NA's                                  :14662  
                                    a24r18        a24rsc1         a24rsc11    

 item not applicable                   :    0   Min.   :1.000   Min.   :0.0
0  
 business & management (general)       :    0   1st Qu.:1.000   1st 
Qu.:0.000  
 enterprises                           :    0   Median :1.000   Median 
:0.000  
 management skills systems & techniques:    0   Mean   :1.098   Mean   
:0.242  
 human resources management            :    0   3rd Qu.:1.000   3rd Qu.:0.0
0  
 (Other)                               :    0   Max.   :3.000   Max.   :1.0

 NA's                                  :14662   NA's   :14398   NA's   :143

    a24rsc12        a24rsc13        a24rsc18        a24rsc19    
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0      
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0      
 Median :0.000   Median :0.000   Median :1.000   Median :0      
 Mean   :0.205   Mean   :0.068   Mean   :0.583   Mean   :0      
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0      
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0      
 NA's   :14398   NA's   :14398   NA's   :14398   NA's   :14398  
   a24rsa2                                                         a24r21     
 Length:14662       office and secretarial skills                 :   10  
 Class :character   information technology & computer applications:    7  
 Mode  :character   business & management (general)               :    6  
                    not stated                                    :    3  
                    public administration                         :    1  
                    (Other)                                       :    3  
                    NA's                                          :14632  
                                    a24r22     
 office and secretarial skills         :    1  
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 (Other)                               :    0  
 NA's                                  :14661  
                                    a24r23     
 item not applicable                   :    0  
 business & management (general)       :    0  
 enterprises                           :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24r24     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24r25     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24r26     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24r27     
 item not applicable                    :    0  
 business & management (general)        :    0  
 enterprises                            :    0  
 management skills systems & techniques:    0  
 human resources management             :    0  
 (Other)                                :    0  
 NA's                                   :14662  
                                    a24r28        a24rsc2         a24rsc21

 item not applicable                    :    0   Min.   :1.000   Min.   :0.0
0  
 business & management (general)        :    0   1st Qu.:1.000   1st 
Qu.:0.000  
 enterprises                            :    0   Median :1.000   Median 
:0.000  
 management skills systems & techniques:    0   Mean   :1.036   Mean   
:0.071  
 human resources management             :    0   3rd Qu.:1.000   3rd Qu.:0.0
0  
 (Other)                                :    0   Max.   :2.000   Max.   :1.0

 NA's                                   :14662   NA's   :14634   NA's   :146

    a24rsc22        a24rsc23        a24rsc28        a24rsc29    
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0      
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0      
 Median :0.000   Median :0.000   Median :1.000   Median :0      
 Mean   :0.214   Mean   :0.071   Mean   :0.679   Mean   :0      
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0      
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0      
 NA's   :14634   NA's   :14634   NA's   :14634   NA's   :14634  
   a24rsa3     
```

```
Length:14662
Class :character
Mode  :character




                                                         a24r31
office and secretarial skills                               :    5
business & management (general)                             :    2
information technology & computer applications              :    2
insufficient info                                           :    1
other - includes child development, hairdressing and beauty :    1
(Other)                                                     :    0
NA's                                                        :14651
                                   a24r32
financial management & accounting     :    1
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
(Other)                               :    0
NA's                                  :14661
                                   a24r33
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24r34
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24r35
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24r36
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24r37
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
```

```
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                 a24r38          a24rsc3         a24rsc31

 item not applicable                  :    0   Min.   :1       Min.   :0

 business & management (general)      :    0   1st Qu.:1       1st Qu.:0

 enterprises                          :    0   Median :1       Median :0

 management skills systems & techniques:    0   Mean   :1       Mean   :0

 human resources management           :    0   3rd Qu.:1       3rd Qu.:0

 (Other)                              :    0   Max.   :1       Max.   :0

 NA's                                 :14662   NA's   :14651   NA's   :146

    a24rsc32         a24rsc33         a24rsc38        a24rsa4
 Min.   :0       Min.   :0.000   Min.   :0.000   Length:14662
 1st Qu.:0       1st Qu.:0.000   1st Qu.:1.000   Class :character
 Median :0       Median :0.000   Median :1.000   Mode  :character
 Mean   :0       Mean   :0.182   Mean   :0.818
 3rd Qu.:0       3rd Qu.:0.000   3rd Qu.:1.000
 Max.   :0       Max.   :1.000   Max.   :1.000
 NA's   :14651   NA's   :14651   NA's   :14651
                                         a24r41
 information technology & computer applications:   48
 office and secretarial skills                  :   12
 not stated                                     :    5
 languages & language studies                   :    3
 financial management & accounting              :    2
 (Other)                                        :    4
 NA's                                           :14588
                            a24r42
 financial management & accounting:    1
 languages & language studies     :    1
 item not applicable              :    0
 business & management (general)  :    0
 enterprises                      :    0
 (Other)                          :    0
 NA's                             :14660
                                 a24r43
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                 a24r44
 item not applicable                  :    0
 business & management (general)      :    0
 enterprises                          :    0
 management skills systems & techniques:    0
 human resources management           :    0
 (Other)                              :    0
 NA's                                 :14662
                                 a24r45
 item not applicable                     :    0
 business & management (general)         :    0
 enterprises                             :    0
 management skills systems & techniques:     0
 human resources management              :    0
 (Other)                                 :    0
 NA's                                    :14662
                                    a24r46
 item not applicable                     :    0
 business & management (general)         :    0
 enterprises                             :    0
 management skills systems & techniques:     0
 human resources management              :    0
 (Other)                                 :    0
 NA's                                    :14662
                                    a24r47
 item not applicable                     :    0
 business & management (general)         :    0
 enterprises                             :    0
 management skills systems & techniques:     0
 human resources management              :    0
 (Other)                                 :    0
 NA's                                    :14662
                                    a24r48        a24rsc4         a24rsc41

 item not applicable                     :    0   Min.   :1.000   Min.   :0.0
0
 business & management (general)         :    0   1st Qu.:1.000   1st
Qu.:0.000
 enterprises                             :    0   Median :1.000   Median
:0.000
 management skills systems & techniques:     0   Mean   :1.091   Mean
:0.182
 human resources management              :    0   3rd Qu.:1.000   3rd Qu.:0.0
0
 (Other)                                 :    0   Max.   :2.000   Max.   :1.0

 NA's                                    :14662   NA's   :14596   NA's   :145

    a24rsc42        a24rsc43        a24rsc48        a24rsc49
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0
 Median :0.000   Median :0.000   Median :1.000   Median :0
 Mean   :0.136   Mean   :0.045   Mean   :0.727   Mean   :0
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0
 NA's   :14596   NA's   :14596   NA's   :14596   NA's   :14596
                                                         a24oq1
 other qualification: band not known, i.e. all other courses :   84
 other band c n.e.c. at nvq level not stated                 :   37
 professional qualifications (further education codes 501-999:   22
 qualification not stated                                    :   17
 unclear/uncodeable                                          :   14
 (Other)                                                     :   26
 NA's                                                        :14462
                                                         a24ol1
 information technology & computer applications             :   29
 other - includes child development, hairdressing and beauty :   29
 nursing                                                     :   17
 sports studies & combined sports                            :   12
```

```
languages & language studies                                :   11
(Other)                                                     :  102
NA's                                                        :14462
                            a24o12
history                        :    1
science & technology (general) :    1
item not applicable            :    0
business & management (general):    0
enterprises                    :    0
(Other)                        :    0
NA's                           :14660
                                                         a24o13
mathematics                                                 :    1
other - includes child development, hairdressing and beauty :    1
item not applicable                                         :    0
business & management (general)                             :    0
enterprises                                                 :    0
(Other)                                                     :    0
NA's                                                        :14660
                                   a24o14
electrical engineering                :    1
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
(Other)                               :    0
NA's                                  :14661
                                   a24o15
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o16
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o17
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o18                      a24ol1
item not applicable                   :    0   item not applicable:    0
business & management (general)       :    0   level 1            :   10
enterprises                           :    0   level 2            :    9
management skills systems & techniques:    0   level 3            :    7
human resources management            :    0   not sure           :  137
(Other)                               :    0   not answered       :   37
NA's                                  :14662   NA's               :14462
                                                         a24oq2
other qualification: band not known, i.e. all other courses :    7
other qualification: band not known, i.e. all other courses :    7
unclear/uncodeable                                          :    2
nvq (not rsa, btec or c & g) level 4                        :    1
professional qualifications (further education codes 501-999:    1
no qualification                                            :    1
(Other)                                                     :    0
NA's                                                        :14650
                            a24o21
dance                          :    2
business & management (general):    1
office and secretarial skills  :    1
fabric crafts                  :    1
history                        :    1
(Other)                        :    6
NA's                           :14650
                                   a24o22
languages & language studies          :    1
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
(Other)                               :    0
NA's                                  :14661
                                   a24o23
languages & language studies          :    1
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
(Other)                               :    0
NA's                                  :14661
                                   a24o24
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o25
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o26
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o27
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o28                      a24ol2
item not applicable                   :    0   item not applicable:    0
business & management (general)       :    0   level 1            :    0
enterprises                           :    0   level 2            :    1
management skills systems & techniques:    0   level 3            :    0
human resources management            :    0   not sure           :    8
(Other)                               :    0   not answered       :    3
NA's                                  :14662   NA's               :14650
                                                         a24oq3
other qualification: band not known, i.e. all other courses :    3
no qualification                                            :    1
qualification not stated                                    :    1
item not applicable                                         :    0
part 1 gnvq foundation                                      :    0
(Other)                                                     :    0
NA's                                                        :14657
                            a24o31
business & management (general):    1
office and secretarial skills  :    1
dance                          :    1
theatre & dramatic arts        :    1
music performance              :    1
(Other)                        :    0
NA's                           :14657
                                   a24o32
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o33
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o34
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o35
item not applicable                   :    0
business & management (general)       :    0
enterprises                           :    0
management skills systems & techniques:    0
human resources management            :    0
(Other)                               :    0
NA's                                  :14662
                                   a24o36
item not applicable                   :    0
```

```
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24o37
 item not applicable                   :    0
 business & management (general)       :    0
 enterprises                           :    0
 management skills systems & techniques:    0
 human resources management            :    0
 (Other)                               :    0
 NA's                                  :14662
                                    a24o38                          a24ol3
 item not applicable                   :    0   item not applicable:    0
 business & management (general)       :    0   level 1            :    0
 enterprises                           :    0   level 2            :    0
 management skills systems & techniques:    0   level 3            :    0
 human resources management            :    0   not sure           :    3
 (Other)                               :    0   not answered       :    2
 NA's                                  :14662   NA's               :14657
                 a25                            a26            a27
 not answered (9)   :    0   not answered (9)   :   0   sep    :1127
 item not applicable:    0   item not applicable:   0   jul    : 780
 yes                :10232   yes                :7853   aug    : 737
 no                 : 4272   no                 :2364   jun    : 729
 NA's               :  158   NA's               :4445   oct    : 723
                                                        (Other):3757
                                                        NA's   :6809
      a28                a30            a31                        a32
 Min.   :102    1-9        :2119   Min.   : 0.000   not answered (9)   :

 1st Qu.:621    10-24      :2024   1st Qu.: 0.000   item not applicable:
0
 Median :720    100 or more:1442   Median : 4.000   employee           :758

 Mean   :704    25-49      :1289   Mean   : 4.304   self-employee      : 11

 3rd Qu.:792    50-99      : 833   3rd Qu.: 7.000   NA's               :696

 Max.   :998    (Other)    :   0   Max.   :18.000

 NA's   :6818   NA's       :6955
                 a33                           a34             a35_1
 not answered (9)   :   0   not answered (9)   :   0   Min.   :  1.00
 item not applicable:   0   item not applicable:   0   1st Qu.: 25.00
 permanent          :5006   yes                :1133   Median : 40.00
 temporary          :1816   no                 :6181   Mean   : 52.61
 not sure           : 936   not sure           : 458   3rd Qu.: 65.00
 NA's               :6904   NA's               :6890   Max.   :450.00
                                                       NA's   :9489
     a35_2            a35               a36
 Min.   : 12.0   Min.   :  2.77   Min.   : 1.00
 1st Qu.:100.0   1st Qu.: 23.08   1st Qu.: 8.00
 Median :160.0   Median : 36.92   Median :14.00
 Mean   :212.2   Mean   : 48.97   Mean   :20.04
 3rd Qu.:280.0   3rd Qu.: 64.62   3rd Qu.:36.00
 Max.   :900.0   Max.   :207.69   Max.   :80.00
 NA's   :12454   NA's   :12454    NA's   :7112
```

```
                                 a37             a37a_1            a37a_2    
not answered (9)                   :   0   Min.   :  5.00   Min.   : 39.97  

item not applicable                :   0   1st Qu.: 37.12   1st Qu.:140.00  

one job or training place          :7149   Median : 51.50   Median :200.00  
more than one job or training place: 526   Mean   : 67.88   Mean   :232.00  

NA's                               :6987   3rd Qu.: 80.00   3rd Qu.:300.00  
                                           Max.   :300.00   Max.   :666.21  
                                           NA's   :14336    NA's   :14499   
     a37a            a37b                         a38      
Min.   :  9.22   Min.   : 1.00   not answered (9)   :   0  
1st Qu.: 32.31   1st Qu.:12.00   item not applicable:   0  
Median : 46.15   Median :16.00   yes                :1929  
Mean   : 53.54   Mean   :20.66   no                 :4899  
3rd Qu.: 69.23   3rd Qu.:27.00   not sure           : 863  
Max.   :153.74   Max.   :60.00   NA's               :6971  
NA's   :14499    NA's   :14201                             
                             a38_2                         a39a     
not answered (9)                :    0   not answered (9)   :   0  
item not applicable             :    0   item not applicable:   0  
yes                             : 1734   yes                :1166  
no                              :  106   no                 :5828  
i have not received any training:   21   not sure           : 660  
my training has not yet started :   61   NA's               :7008  
NA's                            :12740                             
                        a39b                           a40       
modern apprenticeship (ma):  544   not answered (9)   :    0  
youth training (yt)       :  420   item not applicable:    0  
national trainee (ntr)    :   49   yes                :  955  
unclear                   :   31   no                 :  156  
other                     :   21   NA's               :13551  
(Other)                   :   65                              
NA's                      :13532                              
                     a41                            a42       
not answered (9)       :    0   not answered (9)   :    0  
item not applicable    :    0   item not applicable:    0  
a full-time job        :  998   yes                :   18  
a part-time job        :   58   no                 :   48  
it is not part of a job:   68   NA's               :14596  
NA's                   :13538                              

                 a43                        a44_1     
not answered (9)   :    0   not answered (9)   :   0  
item not applicable:    0   item not applicable:   0  
yes                :  746   yes                :3668  
no                 :  291   no                 :4019  
NA's               :13625   NA's               :6975  


                a44_2                       a45_1     
not answered (9)   :    0   not answered (9)   :   0  
item not applicable:    0   item not applicable:   0  
yes                : 2081   yes                :1174  
no                 : 1564   no                 :6484  
NA's               :11017   NA's               :7004  


                a45_2                       a46             a47
```

```
                   _
not answered (9)   :   0   not answered (9)   :    0   Min.   :1.000
item not applicable:   0   item not applicable:    0   1st Qu.:1.000
yes                : 348   yes                :  661   Median :1.000
no                 :5847   no                 :  474   Mean   :1.049
NA's               :8467   NA's               :13527   3rd Qu.:1.000
                                                       Max.   :3.000
                                                       NA's   :13559
     a471            a472            a473            a474
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
Median :0.000   Median :0.000   Median :0.000   Median :0.000
Mean   :0.334   Mean   :0.032   Mean   :0.129   Mean   :0.336
3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000
Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
NA's   :13559   NA's   :13559   NA's   :13559   NA's   :13559
     a475                       a48_1            a48_2
Min.   :0.000   not answered (9)   :    0   Min.   : 1.000
1st Qu.:0.000   item not applicable:    0   1st Qu.: 4.000
Median :0.000   yes                :  503   Median : 4.000
Mean   :0.218   no                 :  636   Mean   : 3.678
3rd Qu.:0.000   NA's               :13523   3rd Qu.: 4.000
Max.   :1.000                               Max.   :10.000
NA's   :13559                               NA's   :14181
                a49_1           a49_2a          a49_2b
not answered (9)   :    0   Min.   :1.000   Min.   : 1.00
item not applicable:    0   1st Qu.:1.000   1st Qu.: 5.00
yes                :  182   Median :1.000   Median :12.00
no                 :  940   Mean   :2.371   Mean   :13.68
NA's               :13540   3rd Qu.:4.000   3rd Qu.:20.00
                            Max.   :8.000   Max.   :52.00
                            NA's   :14600   NA's   :14556
                 a50                               a51
not answered (9)   :    0   not answered (9)      :    0
item not applicable:    0   item not applicable   :    0
excellent          :  329   too much              :   31
good               :  596   not enough            :  169
fair               :  172   about the right amount:  923
poor               :   29   NA's                  :13539
NA's               :13536
                                       a52                          a52a
not answered (9)                         :   0   not answered (9)   :   0
item not applicable                      :   0   item not applicable:   0
yes full-time work (over 30 hours a week):2294   yes                : 340
yes part-time work                       :5439   no                 :5177
yes an occasional job                    : 776   NA's               :9145
no                                       :5947
NA's                                     : 206
                 a52b                        a52c
not answered (9)   :   0   not answered (9)   :   0
item not applicable:   0   item not applicable:   0
yes                : 313   yes                : 391
no                 :5118   no                 :5081
NA's               :9231   NA's               :9190


                a53_1                      a53_2
not answered (9)   :   0   not answered (9)   :    0
item not applicable:   0   item not applicable:    0
yes                :2011   yes                : 1738
no                 :3799   no                 :  251
```

```
NA's               :8852   NA's               :12673



                                                         a53_3
i am a full time student                                    : 3411
i am pregnant/looking after home/children/family            :   66
i believe there is nothing available                        :   61
other                                                       :   42
waiting to start a new job/government supported training/tra:   31
(Other)                                                     :  141
NA's                                                        :10910
     a54a            a54b                           a55_1           a55_2a
Min.   :  1.0   Min.   :  1.0   not answered (9)   :    0   Min.   :  1.05
1st Qu.:100.0   1st Qu.: 80.0   item not applicable:    0   1st Qu.: 11.00

Median :140.0   Median :100.0   yes                :  670   Median : 12.76

Mean   :154.7   Mean   :118.3   no                 :13429   Mean   : 23.04

3rd Qu.:180.0   3rd Qu.:140.0   NA's               :  563   3rd Qu.: 30.00

Max.   :900.0   Max.   :900.0                               Max.   :100.17
NA's   :12011   NA's   :11863                               NA's   :14253
    a55_2b           a55_2            a56_1              a56_2
Min.   :  1.00   Min.   :  0.50   Length:14662       Length:14662
1st Qu.: 30.00   1st Qu.: 11.00   Class :character   Class :character
Median : 59.20   Median : 15.00   Mode  :character   Mode  :character
Mean   : 60.86   Mean   : 24.83
3rd Qu.: 77.80   3rd Qu.: 34.70
Max.   :200.10   Max.   :100.17
NA's   :14531    NA's   :14122
   a56_3a              a56_3b
Length:14662       Min.   : 1.00
Class :character   1st Qu.: 1.00
Mode  :character   Median : 1.00
                   Mean   : 1.69
                   3rd Qu.: 2.00
                   Max.   :23.00
                   NA's   :3231
                                              a56_4a          a56_4b
other                                            :  452   Min.   : 1.000
grandparent(s)                                   :  381   1st Qu.: 1.000
spouse/partner (including boy/girlfriend, fianc0):  113   Median : 1.000
respondent's own child(ren)                      :   33   Mean   : 1.538
not answered (9)                                 :    0   3rd Qu.: 2.000
(Other)                                          :    0   Max.   :12.000
NA's                                             :13683   NA's   :13632
   a56_5                       a57fa                            a57ma
Length:14662       not answered (9)   :    0   not answered (9)   :   0
Class :character   item not applicable:    0   item not applicable:   0
Mode  :character   yes                :11423   yes                :7734
                   no                 : 2005   no                 :6189
                   NA's               : 1234   NA's               : 739



    a57fb         a57mb                           a57fe
Min.   :100   Min.   :101.0   not answered (9)   :   0
1st Qu.:242   1st Qu.:400.0   item not applicable:   0
Median :532   Median :644.0   yes                :3332
Mean   :571   Mean   :621.5   no                 :9584
```

```
3rd Qu.:889   3rd Qu.:958.0   NA's               :1746  
Max.   :998   Max.   :998.0                             
NA's   :7     NA's   :1                                 
                a57me                       a57ff      
not answered (9)   :    0   not answered (9)   :   0  
item not applicable:    0   item not applicable:   0  
yes                : 1297   yes                :3320  
no                 :11577   no                 :6510  
NA's               : 1788   not sure           :3220  
                            NA's               :1612  
                                                      
                a57mf                      a57fg     
not answered (9)   :   0   not answered (9)   :   0  
item not applicable:   0   item not applicable:   0  
yes                :3204   yes                :2629  
no                 :7297   no                 :7787  
not sure           :2838   not sure           :2629  
NA's               :1323   NA's               :1617  
                                                      
                a57mg                       a58       
not answered (9)   :   0   white              :12993  
item not applicable:   0   indian             :  436  
yes                :1937   pakistani          :  280  
no                 :8914   mixed ethnic origin:  126  
not sure           :2458   bangladeshi        :  112  
NA's               :1353   (Other)            :  544  
                           NA's               :  171  
                 a59       
not answered (9)   :    0  
item not applicable:    0  
yes                :  577  
no                 :13865  
NA's               :  220  



                                                          a60       
owned by your parents or yourself                           :11671  
rented from the council                                     : 1736  
rented privately                                            :  433  
rented from a housing association                           :  339  
house/accommodation comes with the job (including police/arm:   93  
(Other)                                                     :  114  
NA's                                                        :  276  
                            a61             a62              a621        
never                        :9414   Min.   :1.000   Min.   :0.00000  
for the odd day or lesson    :3710   1st Qu.:1.000   1st Qu.:0.00000  
for particular days or lessons: 795   Median :1.000   Median :0.00000  
for several days at a time    : 284   Mean   :1.002   Mean   :0.00648  
for weeks at a time           : 246   3rd Qu.:1.000   3rd Qu.:0.00000  
(Other)                       :   0   Max.   :2.000   Max.   :1.00000  
NA's                          : 213   NA's   :146     NA's   :146      
     a622            a623                            a63a     
Min.   :0.000   Min.   :0.000   not answered (9)   :   0  
1st Qu.:0.000   1st Qu.:1.000   item not applicable:   0  
Median :0.000   Median :1.000   agree              :3815  
Mean   :0.058   Mean   :0.938   disagree           :4583  
3rd Qu.:0.000   3rd Qu.:1.000   don't know         :6100  
Max.   :1.000   Max.   :1.000   NA's               : 164  
NA's   :146     NA's   :146                               
                 a63b                        a63c     
```

```
not answered (9)   :   0   not answered (9)   :    0
item not applicable:   0   item not applicable:    0
agree              :5762   agree              : 1004
disagree           :5733   disagree           :12734
don't know         :2967   don't know         :  749
NA's               : 200   NA's               :  175

                 a63d                        a63e          change
not answered (9)   :    0   not answered (9)   :   0   Min.   :0
item not applicable:    0   item not applicable:   0   1st Qu.:0
agree              :10738   agree              :9402   Median :0
disagree           : 1885   disagree           :3369   Mean   :0
don't know         : 1857   don't know         :1707   3rd Qu.:0
NA's               :  182   NA's               : 184   Max.   :0

   change1          change2          change3          change4
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
Median :1.0000   Median :0.0000   Median :1.0000   Median :1.0000
Mean   :0.6865   Mean   :0.3265   Mean   :0.6735   Mean   :0.8269
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

    s1wexp     s1wexp1          s1wexp2          s1wexp3
Min.   :0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000
Median :0   Median :1.0000   Median :1.0000   Median :0.0000
Mean   :0   Mean   :0.9437   Mean   :0.9134   Mean   :0.1851
3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
Max.   :0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

   s1wexp4          s1wexp5          s1wexp6           s1wexp7
Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000
Median :0.0000   Median :0.0000   Median :0.00000   Median :0.000000
Mean   :0.4935   Mean   :0.2058   Mean   :0.02762   Mean   :0.001364
3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000000
Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000

   s1wexp8           s1wexp9            s1nra       s1nra1
Min.   :0.00000   Min.   :0.00000   Min.   :0   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0   1st Qu.:1.0000
Median :0.00000   Median :0.00000   Median :0   Median :1.0000
Mean   :0.05627   Mean   :0.08655   Mean   :0   Mean   :0.8776
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:1.0000
Max.   :1.00000   Max.   :1.00000   Max.   :0   Max.   :1.0000

    s1nra2          s1nra3           s1nra4           s1nra5
Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
Median :1.000   Median :0.0000   Median :0.0000   Median :0.00000
Mean   :0.627   Mean   :0.4314   Mean   :0.1711   Mean   :0.09637
3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000

    s1nra6            s1nra7            s1csown       s1csown1
Min.   :0.00000   Min.   :0.000000   Min.   :0     Min.   :0.000
1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0     1st Qu.:1.000
Median :0.00000   Median :0.000000   Median :0     Median :1.000
Mean   :0.01971   Mean   :0.006275   Mean   :0     Mean   :0.953
3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0     3rd Qu.:1.000
```

```
3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0     3rd Qu.:1.000  
Max.   :1.00000   Max.   :1.000000   Max.   :0     Max.   :1.000  
                                     NA's   :701   NA's   :701    
   s1csown2         s1csown3         s1csown4         s1csown5    
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000  
1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000  
Median :1.0000   Median :1.0000   Median :0.0000   Median :0.000  
Mean   :0.7984   Mean   :0.7973   Mean   :0.4499   Mean   :0.047  
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000  
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000  
NA's   :701      NA's   :701      NA's   :701      NA's   :701    
    s1csgp       s1csgp1          s1csgp2          s1csgp3      
Min.   :0     Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
1st Qu.:0     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
Median :0     Median :1.0000   Median :1.0000   Median :1.0000  
Mean   :0     Mean   :0.6324   Mean   :0.5189   Mean   :0.5123  
3rd Qu.:0     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
Max.   :0     Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
NA's   :701   NA's   :701      NA's   :701      NA's   :701     
   s1csgp4          s1csgp5           s1car       s1car1      
Min.   :0.0000   Min.   :0.0000   Min.   :0   Min.   :0.0000  
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0   1st Qu.:1.0000  
Median :0.0000   Median :0.0000   Median :0   Median :1.0000  
Mean   :0.2583   Mean   :0.3676   Mean   :0   Mean   :0.8558  
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0   3rd Qu.:1.0000  
Max.   :1.0000   Max.   :1.0000   Max.   :0   Max.   :1.0000  
NA's   :701      NA's   :701                                  
    s1car2            s1car3           s1car4          s1car5       
Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   :0.00000  
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00000  
Median :0.00000   Median :0.0000   Median :0.000   Median :0.00000  
Mean   :0.06056   Mean   :0.4709   Mean   :0.263   Mean   :0.05995  
3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.00000  
Max.   :1.00000   Max.   :1.0000   Max.   :1.000   Max.   :1.00000  
                                                                    
    s1car6            s1car7            s1car9       
Min.   :0.000000   Min.   :0.0000   Min.   :0.00000  
1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000  
Median :0.000000   Median :0.0000   Median :0.00000  
Mean   :0.001432   Mean   :0.1343   Mean   :0.00989  
3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.00000  
Max.   :1.000000   Max.   :1.0000   Max.   :1.00000  
                                                     
                    s1sch                         s1acqu         s1qstd 
state school (other)   :11114   5+ gcses!at a*-c     :8415   Min.   :0  
state school (gm)      : 2495   1-4 gcses!at a*-c    :3709   1st Qu.:0  
independent school     : 1053   5+ gcses!at d-g      :1451   Median :0  
not answered (9)       :    0   none!reported        : 757   Mean   :0  
schedule not obtained  :    0   1-4 gcses!at d-g     : 330   3rd Qu.:0  
schedule not applicable:    0   schedule not obtained:   0   Max.   :0  
(Other)                :    0   (Other)              :   0              
   s1qstd01         s1qstd02         s1qstd03         s1qstd04     
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000  
Mean   :0.4772   Mean   :0.1382   Mean   :0.1529   Mean   :0.1087  
3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
                                                                   
   s1qstd05          s1qstd06          s1qstd07         s1qstd08       
Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000000  
```

```
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000000  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000000  
 Median :0.00000   Median :0.00000   Median :0.0000   Median :0.000000  
 Mean   :0.01241   Mean   :0.01446   Mean   :0.1075   Mean   :0.008116  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.000000  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.000000  
                                                                          
    s1qstd09           s1qstd10     
 Min.   :0.000000   Min.   :0.0000  
 1st Qu.:0.000000   1st Qu.:0.0000  
 Median :0.000000   Median :0.0000  
 Mean   :0.001364   Mean   :0.1768  
 3rd Qu.:0.000000   3rd Qu.:0.0000  
 Max.   :1.000000   Max.   :1.0000  
                                    

                                         s1loced    
 state school                                :4884  
 cfe (state system)                          :3198  
 sixth form college (state system)           :1617  
 independent/private school                  : 860  
 institution not stated (difft. from year 11): 245  
 (Other)                                     : 202  
 NA's                                        :3656  
                       s1act1          s1wtrn        s1wtrn1     
 ft education:             :10901   Min.   :0      Min.   :0.000  
 gst:                      : 1398   1st Qu.:0      1st Qu.:0.000  
 ft job:                   : 1182   Median :0      Median :0.000  
 out of work / unemployed::  604   Mean   :0      Mean   :0.175  
 pt job:                   :  332   3rd Qu.:0      3rd Qu.:0.000  
 doing something else:     :  154   Max.   :0      Max.   :1.000  
 (Other)                   :   91   NA's   :6693   NA's   :6693   
    s1wtrn2         s1wtrn3         s1wtrn4         s1wtrn5     
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :0.000   Median :0.000   Median :0.000  
 Mean   :0.142   Mean   :0.261   Mean   :0.083   Mean   :0.063  
 3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000  
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000  
 NA's   :6693    NA's   :6693    NA's   :6693    NA's   :6693    
    s1wtrn6         s1wtrn7                                       s1gst      
 Min.   :0.000   Min.   :0.000   modern apprenticeship (ma)     :  535  
 1st Qu.:0.000   1st Qu.:0.000   youth training (yt)            :  410  
 Median :0.000   Median :1.000   didn't answer question on type :  333  
 Mean   :0.023   Mean   :0.627   other training (write in below):   72  
 3rd Qu.:0.000   3rd Qu.:1.000   national traineeship (ntr)     :   48  
 Max.   :1.000   Max.   :1.000   (Other)                        :    0  
 NA's   :6693    NA's   :6693    NA's                           :13264  
      pseg       pseg1            pseg2            pseg3            pseg4       

 Min.   :0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00
0  
 1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st 
Qu.:0.000  
 Median :0   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0
00  
 Mean   :0   Mean   :0.2299   Mean   :0.2064   Mean   :0.3141   Mean   :0.10
8  
 3rd Qu.:0   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd 
Qu.:0.000  
 Max.   :0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00
0  
```

```
     pseg5              pseg6               pseg7             pseg8       
 Min.   :0.00000   Min.   :0.000000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.00000   Median :0.000000   Median :0.0000   Median :0.0000  
 Mean   :0.03581   Mean   :0.003819   Mean   :0.1019   Mean   :0.4363  
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.00000   Max.   :1.000000   Max.   :1.0000   Max.   :1.0000  

     pseg9                      s1acqe                                   s1emplo    
 Min.   :0.000   5+ gcses at a*-c :8465   in job, but unknown if ft/pt:5169  
 1st Qu.:0.000   1-4 gcses at a*-c:3676   gst                         :1398  

 Median :0.000   5+ gcses at d-g  :1448   ft job                      :1182  

 Mean   :0.458   none reported    : 745   pt job                      : 332  

 3rd Qu.:1.000   1-4 gcses at d-g : 328   not answered (9)            :   0  

 Max.   :1.000   not answered (9) :   0   (Other)                     :   0  

                 (Other)          :   0   NA's                        :6581  
    s1ed_tr     s1ed_tr1          s1ed_tr2           s1ed_tr3       
 Min.   :0   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000  
 1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000  
 Median :0   Median :1.0000   Median :0.00000   Median :0.000000  
 Mean   :0   Mean   :0.7435   Mean   :0.09535   Mean   :0.001569  
 3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.000000  
 Max.   :0   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000  

    s1ed_tr4          s1ed_tr5           s1ed_tr6          s1ed_tr7     
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000  
 Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000  
 Mean   :0.03424   Mean   :0.01514   Mean   :0.8882   Mean   :0.1118  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  

                  s1ecact                          s1eth      
 ilo employed         :9146   white                :12993  
 econ. inactive       :3444   asian groups         : 1005  
 ilo unemployed       :1724   black groups         :  260  
 not answered (9)     :   0   mixed/!other groups:  220  
 schedule not obtained:   0   refused/ns           :  184  
 (Other)              :   0   not answered (9)     :    0  
 NA's                 : 348   (Other)              :    0  
                     xq63a                               xq63b     
 not answered (9)        :   0   not answered (9)        :   0  
 schedule not obtained  :   0   schedule not obtained  :   0  
 schedule not applicable:   0   schedule not applicable:   0  
 item not applicable     :   0   item not applicable     :   0  
 agree                   :3815   agree                   :5762  
 disagree                :4583   disagree                :5733  
 NA's                    :6264   NA's                    :3167  
                     xq63c                                xq63d      
 not answered (9)        :    0   not answered (9)        :    0  
 schedule not obtained  :    0   schedule not obtained  :    0  
 schedule not applicable:    0   schedule not applicable:    0  
 item not applicable     :    0   item not applicable     :    0  
 agree                   : 1004   agree                   :10738  
 disagree               :12734   disagree                 : 1885
 NA's                   :  924   NA's                     : 2039
                   xq63e
 not answered (9)       :   0
 schedule not obtained  :   0
 schedule not applicable:   0
 item not applicable    :   0
 agree                  :9402
 disagree               :3369
 NA's                   :1891
                                              sic                    s1ssr
 whole-!sale, retail, hotels, transp-!ort, etc:5042   other south east:3319
 public, educn., health, other comm., etc     :1010   north west      :1778

 manif-!acturing, elect-!ricity, etc          : 594   west midlands   :1595
 private!hh or! unclass-!ifiable              : 403   greater london  :1572

 finance, real eatate, etc                    : 354   south west      :1423

 (Other)                                      : 450   yorks & humber  :1256
 NA's                                         :6809   (Other)         :3719
               s1tec                     s1denom                    s1mret
 kent               :  520   non denomiantional:10457   march 1998      :724

 devon and cornwall:  453   roman catholic    : 1451   april 1998
:5328
 hampshire          :  436   church of england :  943   may 1998:       :174

 essex              :  434   other             :  851   june 1998       : 33
 sussex             :  427   not applicable    :   51   unclear         :  1

 (Other)            :11492   (Other)           :    9   not answered (9):

 NA's               :  900   NA's              :  900   (Other)         :


                s1agej                    s1ages
 not answered (99)  :   0   not answered (99)  :   0
 item not applicable:   0   item not applicable:   0
 16                 :9878   16                 :6081
 17                 :4784   17                 :8579
 unclear            :   0   unclear            :   2



             s1leaua                                                  s1voqu
 cheshire (x)       :  716   none                                        :13261
 dorset (x)         :  490   unknown level                               :  604
 hampshire (x)      :  454   level 1                                     :  406
 essex (x)          :  432   level 2                                     :  280
 hertfordshire      :  335   level 1 (includes part one foundation):   46
 staffordshire (x):  325   level 3                                     :   35
 (Other)            :11910   (Other)                                     :   30
                                                                    s1avqu
 level 2 (gnvq/nvq full award or 5+ gcses at a-c or 1 a-level:8532
 level 1 (gnvq/nvq full award of 4 gcses any grade)          :5053
 level unknown                                               : 498
 below level 1 (nvq/gnvq certain units only, gnvq part i, 1-3: 295
 none                                                        : 246
 level 3 (gnvq/nvq full award or 2+ a-levels)                :  35
 (Other)                                                     :   3
                                                                    s1peta
```

```
not answered (9)                                            :   0
item not applicable                                         :   0
5+ gcses at a*-c/inter. gnvq and 1+ gcses at a*-c/part one i:8832
1-4 gcses at a*-c/inter. gnvq/part one intermediate gnvq    :3579
5+ gcses at d-g/found'n gnvq and 1+ gcse at d-g/part one fou:1311
1-4 gcse at d-g/found'n gnvq/part one found'n gnvq          : 345
none                                                        : 595
    s1a_c            s1d_g                                s1acqno
Min.   : 0.000   Min.   : 0.000   2+ a level (or equiv)    :6584
1st Qu.: 1.000   1st Qu.: 0.000   none/ns                  :6558
Median : 6.000   Median : 2.000   1-4 gcse                 : 911
Mean   : 5.298   Mean   : 2.847   1-1.5  a level (or equiv): 372
3rd Qu.: 9.000   3rd Qu.: 5.000   other academic           : 106
Max.   :13.000   Max.   :12.000   5+ gcse                  :  90
                                  (Other)                  :  41
                                   s1voqno              s1voqe
none                                   :9193   none         :12768
level 2                                :1990   unknown level:  656
level 3                                :1980   level 2      :  576
level 1                                : 634   level 1      :  557
unknown level (full & some units)      : 610   level 3      :   97
level 2 (includes part one itermediate): 197   level 4      :    8
(Other)                                :  58   (Other)      :    0
                                                        s1hiqua
level 2 (gnvq/nvq full award or 5 gcses at a-c or 1 a level):8700
level 1(gnvq/nvq full award or 4 gcses any grade)           :4852
level unknown                                               : 463
less than level 1 (nvq/gnvq levels 1-4 certain units only, g: 289
none                                                        : 246
level 3 (gnvq/nvq full award or 2+ a levels)                : 104
(Other)                                                     :   8
   s1a_cs1          s1d_gs1          s1mce       s1mce1
Min.   : 0.000   Min.   : 0.000   Min.   :0   Min.   :0.000000
1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.:0   1st Qu.:0.000000
Median : 6.000   Median : 2.000   Median :0   Median :0.000000
Mean   : 5.329   Mean   : 2.866   Mean   :0   Mean   :0.003888
3rd Qu.: 9.000   3rd Qu.: 5.000   3rd Qu.:0   3rd Qu.:0.000000
Max.   :13.000   Max.   :14.000   Max.   :0   Max.   :1.000000

    s1mce2             s1mce3            s1me       s1me01
Min.   :0.000000   Min.   :0.0000   Min.   :0   Min.   :0.0000
1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0   1st Qu.:0.0000
Median :0.000000   Median :1.0000   Median :0   Median :1.0000
Mean   :0.005456   Mean   :0.9945   Mean   :0   Mean   :0.5444
3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0   3rd Qu.:1.0000
Max.   :1.000000   Max.   :1.0000   Max.   :0   Max.   :1.0000

    s1me02           s1me03             s1me04             s1me05
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000
Mean   :0.4854   Mean   :0.04495   Mean   :0.01405   Mean   :0.1709
3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000

    s1me06             s1me07             s1me08             s1me09
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
Mean   :0.08321   Mean   :0.05456   Mean   :0.03315   Mean   :0.2847
```

```
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000

    s1me10            s1me11            s1me12          s1vqtp
Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0
Median :0.00000   Median :0.00000   Median :0.000   Median :0
Mean   :0.07939   Mean   :0.07632   Mean   :0.129   Mean   :0
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0
Max.   :1.00000   Max.   :1.00000   Max.   :1.000   Max.   :0

   s1vqtp01           s1vqtp02           s1vqtp03          s1vqtp04
Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.000000   Median :0.000000   Median :0.00000   Median :0.00000
Mean   :0.007093   Mean   :0.003001   Mean   :0.01159   Mean   :0.00798
3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000

   s1vqtp05           s1vqtp06          s1vqtp07           s1vqtp08
Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
Median :0.000000   Median :0.00000   Median :0.000000   Median :0.000000
Mean   :0.003478   Mean   :0.02046   Mean   :0.002046   Mean   :0.003478
3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000

   s1vqtp09           s1vqtp10           s1vqtp11          s1vqtp12
Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000
1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:1.000
Median :0.000000   Median :0.000000   Median :0.00000   Median :1.000
Mean   :0.002455   Mean   :0.001705   Mean   :0.02483   Mean   :0.919
3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:1.000
Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000

    s1gnvq         s1gnvq1         s1gnvq2         s1gnvq3
Min.   :0       Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:0       1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
Median :0       Median :0.000   Median :0.000   Median :0.000
Mean   :0       Mean   :0.075   Mean   :0.442   Mean   :0.387
3rd Qu.:0       3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.000
Max.   :0       Max.   :1.000   Max.   :1.000   Max.   :1.000
NA's   :12232   NA's   :12232   NA's   :12232   NA's   :12232
   s1gnvq4         s1gnvq5         s1gnvq6         s1gnvq7
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
Median :0.000   Median :0.000   Median :0.000   Median :0.000
Mean   :0.016   Mean   :0.049   Mean   :0.011   Mean   :0.027
3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000
Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
NA's   :12232   NA's   :12232   NA's   :12232   NA's   :12232
    s1nvq          s1nvq01         s1nvq02         s1nvq03
Min.   :0       Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:0       1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
Median :0       Median :0.000   Median :1.000   Median :0.000
Mean   :0       Mean   :0.174   Mean   :0.565   Mean   :0.218
3rd Qu.:0       3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:0.000
Max.   :0       Max.   :1.000   Max.   :1.000   Max.   :1.000
NA's   :12869   NA's   :12869   NA's   :12869   NA's   :12869
   s1nvq04         s1nvq05         s1nvq06         s1nvq07
```

```
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
Median :0.000   Median :0.000   Median :0.000   Median :0.000  
Mean   :0.099   Mean   :0.028   Mean   :0.064   Mean   :0.018  
3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000  
Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000  
NA's   :12869   NA's   :12869   NA's   :12869   NA's   :12869  
   s1nvq08         s1nvq09           s1payh          s1payho      
Min.   :0.00    Min.   :0.00    Min.   :  0.02   Min.   :  0.25  
1st Qu.:0.00    1st Qu.:0.00    1st Qu.:  2.16   1st Qu.:  1.50  
Median :0.00    Median :0.00    Median :  2.88   Median :  2.08  
Mean   :0.03    Mean   :0.06    Mean   : 15.87   Mean   : 32.30  
3rd Qu.:0.00    3rd Qu.:0.00    3rd Qu.:  3.50   3rd Qu.:  2.89  
Max.   :1.00    Max.   :1.00    Max.   :997.00   Max.   :997.00  
NA's   :12869   NA's   :12869   NA's   :7318     NA's   :14228   
                  s1apr97                        s1may97     
ft edn                :10588   ft edn                :10381  
not answered          : 1723   not answered          : 1677  
out of work/unemployed:  911   out of work/unemployed:  916  
something else        :  529   something else        :  600  
pt job                :  468   pt job                :  560  
ft job                :  304   ft job                :  364  
(Other)               :  139   (Other)               :  164  
                  s1jun97                       s1jul97    
ft edn                :8904   ft edn                :4813  
not answered          :1552   pt job                :2452  
out of work/unemployed:1117   something else        :2392  
pt job                :1076   out of work/unemployed:1627  
something else        :1000   not answered          :1549  
ft job                : 651   ft job                :1202  
(Other)               : 362   (Other)               : 627  
                  s1aug97                       s1sep97     
ft edn                :3684   ft edn                :10958  
pt job                :2871   gst                   : 1123  
something else        :2699   ft job                :  993  
out of work/unemployed:1610   pt job                :  506  
not answered          :1559   out of work/unemployed:  496  
ft job                :1404   not answered          :  324  
(Other)               : 835   (Other)               :  262  
                  s1oct97                        s1nov97     
ft edn                :11365   ft edn                :11300  
gst                   : 1166   gst                   : 1177  
ft job                : 1016   ft job                : 1076  
out of work/unemployed:  407   out of work/unemployed:  415  
pt job                :  324   pt job                :  333  
not answered          :  260   not answered          :  241  
(Other)               :  124   (Other)               :  120  
                  s1dec97                        s1jan98     
ft edn                :11105   ft edn                :11038  
gst                   : 1185   gst                   : 1221  
ft job                : 1106   ft job                : 1159  
out of work/unemployed:  470   out of work/unemployed:  524  
pt job                :  376   pt job                :  334  
not answered          :  249   not answered          :  235  
(Other)               :  171   (Other)               :  151  
                  s1feb98     
ft edn                :10962  
gst                   : 1241  
ft job                : 1218  
out of work/unemployed:  532  
```

```
 pt job                :  339
 not answered          :  219
 (Other)               :  151
```

Various WARNINGS will appear. Don't panic.

To see the summary of the data use the scroll bar to **scroll down**.

Get a subset of the data with only the variables needed.

In [11]:

```
myvars <- c("serial", "weight", "sex", "s1a_c", "a58", "s1eth", "s1acqe",
"pseg", "pseg1", "pseg2", "pseg3", "pseg4", "pseg5", "pseg6", "pseg7")
mydata.df <- mydata.df[myvars]
```

In [6]:

```
summary(mydata.df)
```

Out[6]:

```
     serial           weight                            sex              s1a_c

 Min.   :200001   Min.   :0.6025   not answered (9)    :   0   Min.   :
0.000
 1st Qu.:206123   1st Qu.:0.7661   item not applicable:   0   1st Qu.: 1.00
0
 Median :211589   Median :0.8779   male                :6889   Median : 6.00

 Mean   :212056   Mean   :1.0000   female              :7773   Mean   : 5.29

 3rd Qu.:217027   3rd Qu.:1.0576                               3rd Qu.: 9.00
 Max.   :231392   Max.   :2.5176                               Max.   :13.00



                    a58                             s1eth
 white                :12993   white                :12993
 indian               :  436   asian groups         : 1005
 pakistani            :  280   black groups         :  260
 mixed ethnic origin:  126   mixed/!other groups:  220
 bangladeshi          :  112   refused/ns           :  184
 (Other)              :  544   not answered (9)     :    0
 NA's                 :  171   (Other)              :    0
                s1acqe            pseg        pseg1             pseg2
 5+ gcses at a*-c :8465   Min.   :0   Min.   :0.0000   Min.   :0.0000
 1-4 gcses at a*-c:3676   1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.0000
 5+ gcses at d-g  :1448   Median :0   Median :0.0000   Median :0.0000
 none reported    : 745   Mean   :0   Mean   :0.2299   Mean   :0.2064
 1-4 gcses at d-g : 328   3rd Qu.:0   3rd Qu.:0.0000   3rd Qu.:0.0000
 not answered (9) :   0   Max.   :0   Max.   :1.0000   Max.   :1.0000
 (Other)          :   0
     pseg3            pseg4             pseg5              pseg6
 Min.   :0.0000   Min.   :0.000   Min.   :0.00000   Min.   :0.000000
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.000000
 Median :0.0000   Median :0.000   Median :0.00000   Median :0.000000  
 Mean   :0.3141   Mean   :0.108   Mean   :0.03581   Mean   :0.003819  
 3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.000000  
 Max.   :1.0000   Max.   :1.000   Max.   :1.00000   Max.   :1.000000  

     pseg7       
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :0.0000  
 Mean   :0.1019  
 3rd Qu.:0.0000  
 Max.   :1.0000  
```

str compactly display the internal structure of an *R* object.

It is a diagnostic function and an alternative to summary.

In [12]:

```
str(mydata.df)
```

```
'data.frame':	14662 obs. of  15 variables:
 $ serial: int  200001 200004 200005 200006 200008 200012 200013 200014 200
019 200022 ...
 $ weight: num  0.875 0.976 0.976 0.976 1.841 ...
 $ sex   : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 4 4 4 3 4
...
 $ s1a_c : int  9 9 9 9 0 1 5 2 1 1 ...
 $ a58   : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 6 11 6 4 10
10 ...
 $ s1eth : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 6 7 6 5 7 7
...
 $ s1acqe: Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 7 6 5 6 6 6
...
 $ pseg  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg1 : int  1 0 0 0 0 0 0 0 0 0 ...
 $ pseg2 : int  0 1 0 1 0 0 0 0 0 0 ...
 $ pseg3 : int  0 0 1 0 0 0 0 1 0 0 ...
 $ pseg4 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7 : int  0 0 0 0 1 1 1 0 1 1 ...
```

View the data in spreadsheet format.

In [13]:

```
head(mydata.df)
```

Out[13]:

	serial	weight	sex	s1a_c	a58	s1eth	s1acqe	pseg	pseg1	pseg2	pseg3	pseg4
							5+					

	serial	weight	sex	s1a_c	a58	s1eth	s1acqe	pseg	pseg1	pseg2	pseg3	pseg4
1	200001	0.87518	female	9	white	white	gcses at a*-c	0	1	0	0	0
2	200004	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	0	0
3	200005	0.97615	male	9	indian	asian groups	5+ gcses at a*-c	0	0	0	1	0
4	200006	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	0	0
5	200008	1.84073	female	0	afro.	black groups	5+ gcses at d-g	0	0	0	0	0
6	200012	0.95928	female	1	chin!ese	asian groups	1-4 gcses at a*-c	0	0	0	0	0

Construct the outcome variable

The binary indicator of 5+ GCSEs A (star) - C will be "s15a_c" .

It is constructed from variable "s1a_c" - the number of GCSEs A (star) - C.

Tabulate the original outcome "s1a_c" number of GCSEs A (star) - C .

In [14]:

```
table(mydata.df$s1a_c)
```

Out[14]:

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13 
2538 1138  936  833  802  788  854 1023 1373 2459 1525  345   45    3 
```

Construct the binary outcome variable "s15a_c" 5+ GCSEs at grades A-C

Following the existing naming convention used in YCS Cohort 9 I have chosen the title "s15a_c" because this is a sweep 1 measure "s1" of 5+ GCSEs at grades A-C "5a_c" hence "s15a_c".

Create the "empty" new field.

In [15]:

```
mydata.df$s15a_c <- NA
table(mydata.df$s15a_c)
```

Out[15]:

```
< table of extent 0 >
```

The new field "$s15a_c" is empty.

Recode the old field into the new one for the specified rows.

In [16]:

```
mydata.df$s15a_c[mydata.df$s1a_c>4] <-1
table(mydata.df$s15a_c)

mydata.df$s15a_c[mydata.df$s1a_c<5] <-0
table(mydata.df$s15a_c)
```

Out[16]:

```
   1
8415
```

Out[16]:

```
   0    1
6247 8415
```

Construct the first variable explanatory variable girls (gender)

The binary indicator of girls (gender) from the existing variable "sex" .

This is a factor. Therefore I will check levels in the original data boys==1 and girls==2.

In [17]:

```
levels(mydata.df$sex)
table (mydata.df$sex)
```

Out[17]:

"not answered (9)" "item not applicable" "male" "female"

Out[17]:

```
   not answered (9)  item not applicable                     male                fer
ale
                  0                    0                     6889                  '
3
```

Create the "empty" new field.

In [18]:

```
# create the new field
mydata.df$girls <- NA
table(mydata.df$girls)
```

Out[18]:

```
< table of extent 0 >
```

The new field "$girls" is empty.

Recode the old field into the new one for the specified rows.

In [19]:

```
mydata.df$girls[mydata.df$sex=="male"] <-0
mydata.df$girls[mydata.df$sex=="female"] <-1
```

In [20]:

```
table(mydata.df$girls)
```

Out[20]:

```
   0    1 
6889 7773 
```

Construct the explanatory variable for ethnicity

Beware this measure is messy!

This is a factor. Therefore I check levels in the original data.

In [21]:

```
levels(mydata.df$a58)
```

Out[21]:

```
"not answered (99)"   "don't know (98)"   "item not applicable"   "white"   "carib."   "afro."
"other black"   "indian"   "pakistani"   "bangladeshi"   "chin!ese"   "other asian"
"any other"   "mixed ethnic origin"   "ref!used"
```

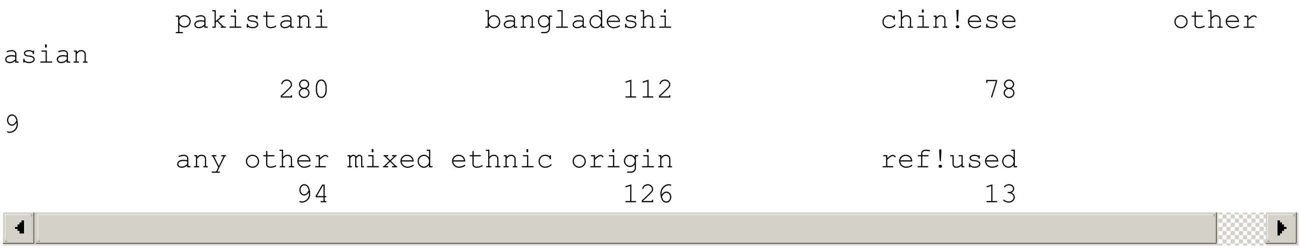
Now I create a table of the ethnicity measure "a58".

In [14]:

```
table(mydata.df$a58)
```

Out[14]:

```
  not answered (99)     don't know (98) item not applicable               wh
ite 
                  0                   0                   0               12
3 
             carib.               afro.         other black              ind
an 
                104                  78                  78                 
6 
```



```
          pakistani         bangladeshi              chin!ese          other
asian
                280                 112                    78
9
          any other mixed ethnic origin              ref!used
                 94                 126                    13
```

These are the dummies that are required for the model in Connolly (2006, p.20)

Chinese
Indian
White
Bangladeshi
Pakistani

Strangely, the "Other" category is not in the model!

Ethnic categories required for the analysis.

These are the categories developed and used in Connolly (2006) Table 1 (p.7)

White Indian Pakistani Black Bangladeshi Chinese Other

Here are the labels and codes in the YCS dataset

- 1 "white"
- 2 "carib." "afro." "other black"
- 3 "indian"
- 4 "pakistani"
- 5 "bangladeshi"
- 6 "chin!ese"
- 7 "other asian"
- 10 "any other"
- 97 "mixed ethnic origin"
- . "ref!used"

Here are my estimates of the number is each ethnic category used in Connolly (2006) Table 1 (p.7)

1	White	12993
2	Black	260
3	Indian	436
4	Pakistani	280
5	Bangladeshi	112
6	Chinese	78
7	Other	503

Create the new field "ethnic1".

Everyone is placed into category 7.

I then recode the new field "ethnic1" with values from "a58" .

There is an explanation of this unorthodox approach below...

In [23]:

```
# create the new field,

mydata.df$ethnic1 <- 7

# everyone is placed into category 7.

# recode the new field with values from the old field.
mydata.df$ethnic1[mydata.df$a58=="white"] <-1
mydata.df$ethnic1[mydata.df$a58=="carib."] <-2
mydata.df$ethnic1[mydata.df$a58=="afro."] <-2
mydata.df$ethnic1[mydata.df$a58=="other black"] <-2
mydata.df$ethnic1[mydata.df$a58=="indian"] <-3
mydata.df$ethnic1[mydata.df$a58=="pakistani"] <-4
mydata.df$ethnic1[mydata.df$a58=="bangladeshi"] <-5
mydata.df$ethnic1[mydata.df$a58=="chin!ese"] <-6
mydata.df$ethnic1[mydata.df$a58=="other asian"] <-7
mydata.df$ethnic1[mydata.df$a58=="any other"] <-7
mydata.df$ethnic1[mydata.df$a58=="mixed ethnic origin"] <-7
mydata.df$ethnic1[mydata.df$a58=="ref!used"] <-7
```

There appears to be a quirk in the labelling of the missing values "." in the Stata file.

I have got around this by forcing these cases into category 7 when I created the new field
i.e. *mydata.df$ethnic1 <- 7*

Create a table of the new "ethnic1" variable.

In [24]:

```
table(mydata.df$ethnic1)
```

Out[24]:

```
    1     2     3     4     5     6     7 
12993   260   436   280   112    78   503 
```

This might not be the neatest solution! But the obstactle has been overcome.

In [25]:

```
# Just to check the variable again.

table(mydata.df$ethnic1)
```

Out[25]:

```
    1     2     3     4     5     6     7 
12993   260   436   280   112    78   503 
```

In [26]:

```
# Double check ethnic1 is not a factor

levels(mydata.df$ethnic1)
```

```
Out[26]:

NULL
```

Construct a series of dummy variables for ethnicity

I have chosen to construct a each variable manually, in order to double check it.

White pupils.

```
In [27]:
```

```
mydata.df$white <-0
table(mydata.df$white)
mydata.df$white[mydata.df$ethnic1=="1"] <-1
table(mydata.df$white)
```

```
Out[27]:

    0 
14662 
```

```
Out[27]:

    0     1 
 1669 12993 
```

Black pupils.

```
In [28]:
```

```
mydata.df$black <-0
table(mydata.df$black)
mydata.df$black[mydata.df$ethnic1=="2"] <-1
table(mydata.df$black)
```

```
Out[28]:

    0 
14662 
```

```
Out[28]:

    0     1 
14402   260 
```

Indian pupils

```
In [29]:
```

```
mydata.df$indian <-0
table(mydata.df$indian)
mydata.df$indian[mydata.df$ethnic1=="3"] <-1
mydata.df$indian[mydata.df$ethnic1=="3"] <-1
table(mydata.df$indian)
```

Out[29]:

```
    0 
14662 
```

Out[29]:

```
    0     1 
14226   436 
```

Pakistani pupils.

In [30]:

```
mydata.df$pakistani <-0
table(mydata.df$pakistani)
mydata.df$pakistani[mydata.df$ethnic1=="4"] <-1
table(mydata.df$pakistani)
```

Out[30]:

```
    0 
14662 
```

Out[30]:

```
    0     1 
14382   280 
```

Bangladeshi pupils.

In [31]:

```
mydata.df$bangladeshi <-0
table(mydata.df$bangladeshi)
mydata.df$bangladeshi[mydata.df$ethnic1=="5"] <-1
table(mydata.df$bangladeshi)
```

Out[31]:

```
    0 
14662 
```

Out[31]:

```
    0     1 
14550   112 
```

Chinese pupils.

In [32]:

```
mydata.df$chinese <-0
table(mydata.df$chinese)
table(mydata.df$chinese)
mydata.df$chinese[mydata.df$ethnic1=="6"] <-1
table(mydata.df$chinese)
```

Out[32]:

```
    0 
14662 
```

Out[32]:

```
    0     1 
14584    78 
```

Other pupils.

In [33]:

```
mydata.df$other <-0
table(mydata.df$other)
mydata.df$other[mydata.df$ethnic1=="7"] <-1
table(mydata.df$other)
```

Out[33]:

```
    0 
14662 
```

Out[33]:

```
    0     1 
14159   503 
```

The block of dummy variables representing ethnicity have been constructed.

Now I perform a brief test.

Here is a table of the outcome variable 5+ GCSEs at grades A - C.

In [34]:

```
table(mydata.df$s15a_c)
```

Out[34]:

```
   0    1 
6247 8415 
```

Here is a table of ethnicity.

In [35]:

```
table(mydata.df$ethnic1)
```

Out[35]:

```
    1     2     3     4     5     6     7 
12993   260   436   280   112    78   503 
```

Here is a table of school GCSE outcome by ethnicity

In [36]:

```
mytable <- table(mydata.df$ethnic1, mydata.df$s15a_c) # A will be rows, B w
ill be columns
mytable # print table
```

Out[36]:

```
       0    1
  1 5433 7560
  2  158  102
  3  160  276
  4  172  108
  5   64   48
  6   20   58
  7  240  263
```

There results look plausible and I am happy that the measures are behaving themselves.

Construct the explanatory variable for social class

Beware this a bit messy!

The variables pseg1 - pseg7 are social class dummies.

I would like these variables to have names that are more "human-eye-readable".

Here is the first social class dummy "pseg1" which is the Professional/Managerial social class.

In [43]:

```
table(mydata.df$pseg1)
```

Out[43]:

```
< table of extent 0 >
```

Here we will be using the reshape library.

Make sure that it has been installed make sure that it has been installed

The *R* code is
install.packages ("reshape")

In [42]:

```
library(reshape)
```

Various WARNINGS might appear. Don't panic.

Here is the code to *rename* "pseg1" as "prof_man" i.e.

In [ ]:

```
mydata.df <- rename(mydata.df, c(pseg1="prof_man"))
```

Now take a look at the "renamed" variable.

In [44]:

```
table(mydata.df$prof_man)
```

Out[44]:

```

    0     1 
11291  3371 
```

I now rename pseg2 - pseg4.

In [46]:

```
mydata.df <- rename(mydata.df, c(pseg2="o_non_man"))
table(mydata.df$o_non_man)
```

Out[46]:

```

    0     1 
11636  3026 
```

In [85]:

```
mydata.df <- rename(mydata.df, c(pseg3="skilled_man"))
table(mydata.df$skilled_man)
```

Out[85]:

```

    0     1 
10056  4606 
```

In [84]:

```
mydata.df <- rename(mydata.df, c(pseg4="semi_skilled"))
table(mydata.df$semi_skilled)
```

Out[84]:

```

    0     1 
13078  1584 
```

The dataset has now been 'wrangled' or 'enabled' and should be in reasonable shape to test.

In the next stage I test the data and ultimately I will try to duplicate the model in Connelly (2006, p.20).

Now I save the wrangled data frame in a file called **ycs9sw1.rda**.

In [87]:

```
save(mydata.df,file="C:/Users/Vernon/OneDrive - University of Edinburgh/Doc
uments/ycs_9_2017/ycs9sw1.rda")
```

List the objects in my workspace.

In [88]:

```
ls()
```

Out[88]:

```
"mydata.df"   "mydesign1"   "small.w"
```

Now I am going to remove "rm" these objects.

In [89]:

```
rm ("mydata.df", "mytable", "myvars")
ls()
```

```
Warning message:
In rm("mydata.df", "mytable", "myvars"): object 'mytable' not foundWarning
message:
In rm("mydata.df", "mytable", "myvars"): object 'myvars' not found
```

Out[89]:

```
"mydesign1"   "small.w"
```

Data Test

In this section I undertake a small series of exploratory data analysis tasks to check the data with the published results in Connolly (2006).

Re-loading the data frame from the saved file.

In [92]:

```
load("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1.rda")
ls()
```

Out[92]:

```
"mydata.df"
```

Now I set up the survey desing of the YCS 9.

Within an object called "small.w" I specify the design.

The "ids" are the identification for each case i.e. "serial".
The data are "mydata.df".
The survey weights are "weight".

In [93]:

```
small.w <- svydesign(ids = ~serial, data = mydata.df, weights = ~weight)
```

Now I attempt to check the values of my variables against the values (*n*) and proportions reported in Connolly (2006, p.7, Table 1).

Girls.

In [96]:

```
table(svytable(~girls, design = small.w))
prop.table(svytable(~girls, design = small.w))
```

Out[96]:

```
7268.86727  7393.1375
         1          1
```

Out[96]:

```
girls
        0         1
0.5042378 0.4957622
```

These results are correct. Checked with Connolly (2006, p.7, Table 1).

Ethnicity.

In [97]:

```
table(svytable(~ethnic1, design = small.w))
prop.table(svytable(~ethnic1, design = small.w))
```

Out[97]:

```
   74.05745    122.49261    297.11676    311.67752     437.4739    525.13248
          1            1            1            1            1            1
12894.05405
          1
```

Out[97]:

```
ethnic1
          1           2           3           4           5           6
0.879419578 0.020264402 0.029837250 0.021257497 0.008354424 0.005050977
          7
0.035815872
```

These results are correct. Checked with Connolly (2006, p.7, Table 1).

Remember that the ordering in Connolly (2006, p.7, Table 1) is not the same as in the logit model Connolly (2006, p.20).

Social class.

Here I remind myself of the variable names.

I use *str* which is a compact display of the "structure" of an arbitrary *R* object.

In [98]:

```
str(mydata.df)
```

```
'data.frame':	14662 obs. of  25 variables:
 $ serial      : int  200001 200004 200005 200006 200008 200012 200013 2000
14 200019 200022 ...
 $ weight      : num  0.875 0.976 0.976 0.976 1.841 ...
 $ sex         : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 4 4 4 
3 4 ...
 $ s1a_c       : int  9 9 9 9 0 1 5 2 1 1 ...
 $ a58         : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 6 11 6 
4 10 10 ...
 $ s1eth       : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 6 7 6 5 
7 7 ...
 $ s1acqe      : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 7 6 5 6 
6 6 ...
 $ pseg        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 0 0 0 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 0 0 ...
 $ skilled_man : int  0 0 1 0 0 0 0 1 0 0 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7       : int  0 0 0 0 1 1 1 0 1 1 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 0 0 ...
 $ girls       : num  1 0 0 0 1 1 1 1 0 1 ...
 $ ethnic1     : num  1 1 3 1 2 6 2 1 5 5 ...
 $ white       : num  1 1 0 1 0 0 0 1 0 0 ...
 $ black       : num  0 0 0 0 1 0 1 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 1 1 ...
 $ chinese     : num  0 0 0 0 0 1 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
```

In [105]:

```
print("prof_man")
table(svytable(~prof_man , design = small.w))
prop.table(svytable(~prof_man , design = small.w))

print("o_non_man")
table(svytable(~o_non_man , design = small.w))
```

```
prop.table(svytable(~o_non_man , design = small.w))

print("skilled_man")
table(svytable(~skilled_man, design = small.w))
prop.table(svytable(~skilled_man , design = small.w))

print("semi_skilled")
table(svytable(~semi_skilled , design = small.w))
prop.table(svytable(~semi_skilled , design = small.w))
```

```
[1] "prof_man"
```

Out[105]:

```
 3048.57466 11613.43011 
          1           1 
```

Out[105]:

```
prof_man
        0         1 
0.7920765 0.2079235 
```

```
[1] "o_non_man"
```

Out[105]:

```
  2829.6733 11832.33147 
          1           1 
```

Out[105]:

```
o_non_man
        0         1 
0.8070064 0.1929936 
```

```
[1] "skilled_man"
```

Out[105]:

```
4697.93136 9964.07341 
         1          1 
```

Out[105]:

```
skilled_man
        0         1 
0.6795847 0.3204153 
```

```
[1] "semi_skilled"
```

Out[105]:

```
 1702.20359 12959.80118 
          1           1 
```

Out[105]:

```
semi_skilled
        0         1 
0.8839038 0.1160962 
```

These results are correct.

Checked with Connolly (2006, p.7, Table 1).

Remember that the categories used in the logit model Connolly (2006, p.20). are not the same as in Connolly (2006, p.7, Table 1).

Connolly (2006, p.7, Table 1).

In this section I have undertaken a small series of exploratory data analysis tasks to check the data with the published results in Connolly (2006).

I am confident that the data are in good shape ready for duplicating the logit model.

Data Analysis

Duplicating the Connelly (2006) Model Results in *R*

Table 5 p.20 Connolly (2006).

<span style="color:red">Beware</span> if you skipped to this section then make sure that you have the correct data frame (i.e. the data file "ycs9sw1.rda").

Re-loading the data frame from the saved file requries this *R* code

_load("C:/Users/Vernon/OneDrive - University of Edinburgh/Documents/ycs_9*2017/ycs9sw1.rda")
ls()*

It is currently in the markdown cell below.

You might also require the *R* libraries.

Setting up the survey design of the YCS data.

In [3]:

```
library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)
load("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1.rda")
ls()
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package:
grid
Loading required package: Matrix
Loading required package: survival
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns
```

Out[3]:

"mydata.df" "mydata4.df"

In [2]:

```
mydesign1 <- svydesign(id = ~serial,data = mydata.df, weight = ~weight)
```

This is a *svy* (i.e.survey) logit regression model.
The outcome variable is "s15a_c" - 5+ GCSEs at grades A - C .

The explanatory variables are

> girls
> ethnicity (represented by a block of dummy variables)

> ethnicity (represented by a block of dummy variables)
> social class (represented by a block of dummy variables).

In [3]:

```
model1<-svyglm (s15a_c ~ girls + chinese + indian +
                 white + bangladeshi + pakistani +
                 pakistani + prof_man + o_non_man +
                 skilled_man + semi_skilled, design=mydesign1, data = mydata
.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

There might be a warning message because we are modelling survey (i.e. weighted) data. Don't panic.

Summary of the model results.

In [4]:

```
summary (model1)
```

Out[4]:

```
Call:
svyglm(formula = s15a_c ~ girls + chinese + indian + white +
    bangladeshi + pakistani + pakistani + prof_man + o_non_man +
    skilled_man + semi_skilled, design = mydesign1, data = mydata.df,
    family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata.df, weight = ~weight)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.58272    0.09190 -17.223  < 2e-16 ***
girls         0.39532    0.03663  10.791  < 2e-16 ***
chinese       1.34282    0.29745   4.514 6.40e-06 ***
indian        0.60915    0.13734   4.435 9.26e-06 ***
white         0.16152    0.08575   1.884   0.0596 .
bangladeshi   0.24018    0.21983   1.093   0.2746
pakistani    -0.06046    0.15696  -0.385   0.7001
prof_man      2.04847    0.06612  30.980  < 2e-16 ***
o_non_man     1.62986    0.06503  25.062  < 2e-16 ***
skilled_man   0.79762    0.05867  13.595  < 2e-16 ***
semi_skilled  0.43251    0.07230   5.982 2.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000384)

Number of Fisher Scoring iterations: 4
```

These results are not the same as the results presented in Table 5, p.20 Connolly (2006).

The difference may not be immediately apparent.

In the published work Connelly (2006)

a) the pupils in ethnic category "other" are dropped from the analysis

b) the pupils in class categories "other" and "unclassified" are dropped from the analysis.

Here we subset ethnic categories 1 - 6.

In [8]:

```
table(mydata.df$ethnic1)
```

Out[8]:

```
    1     2     3     4     5     6     7 
12993   260   436   280   112    78   503 
```

Here we subset ethnic categories 1 - 6.

In [7]:

```
mydata2.df <- subset(mydata.df, ethnic1!=7)
table(mydata.df$ethnic1)
table(mydata2.df$ethnic1)
```

Out[7]:

```
    1     2     3     4     5     6     7 
12993   260   436   280   112    78   503 
```

Out[7]:

```
    1     2     3     4     5     6 
12993   260   436   280   112    78 
```

Here we subset pupils in class categories pseg1 - pseg5.

In [8]:

```
table(mydata2.df$pseg6)
table(mydata2.df$pseg7)
```

Out[8]:

```
    0     1 
14106    53 
```

Out[8]:

```
    0     1 
12842  1317 
```

In [9]:

In [9]:

```
mydata3.df <- subset(mydata2.df, pseg6!=1)
table(mydata.df$pseg6)
table(mydata3.df$pseg6)
```

Out[9]:

```
    0     1 
14606    56 
```

Out[9]:

```
    0 
14106 
```

In [10]:

```
mydata4.df <- subset(mydata3.df, pseg7!=1)
table(mydata.df$pseg7)
table(mydata4.df$pseg7)
```

Out[10]:

```
    0     1 
13168  1494 
```

Out[10]:

```
    0 
12789 
```

The dataset should be in shape now for estimating the (survey) logit model.

Beware you might need to reset the design.

In [19]:

```
mydesign2 <- svydesign(id = ~serial,data = mydata4.df, weight = ~weight)
```

In [22]:

```
model2<-svyglm (s15a_c ~ girls + chinese + indian +
                 white + bangladeshi + pakistani +
                 pakistani + prof_man + o_non_man +
                 skilled_man + semi_skilled, design=mydesign2, data = mydata
4.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

There might be a warning message because we are modelling survey (i.e. weighted) data. Don't panic.

Summary of the model results.

In [23]:

In [23]:

```
summary (model2)
```

Out[23]:

```
Call:
svyglm(formula = s15a_c ~ girls + chinese + indian + white + 
    bangladeshi + pakistani + pakistani + prof_man + o_non_man + 
    skilled_man + semi_skilled, design = mydesign2, data = mydata4.df, 
    family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.20829    0.19802 -11.152  < 2e-16 ***
girls         0.40456    0.03926  10.305  < 2e-16 ***
chinese       2.00231    0.37734   5.306 1.14e-07 ***
indian        1.06584    0.20829   5.117 3.15e-07 ***
white         0.64314    0.17118   3.757 0.000173 ***
bangladeshi   0.76616    0.34486   2.222 0.026323 *  
pakistani     0.53136    0.24503   2.169 0.030135 *  
prof_man      2.19209    0.10863  20.179  < 2e-16 ***
o_non_man     1.77251    0.10793  16.423  < 2e-16 ***
skilled_man   0.93217    0.10411   8.954  < 2e-16 ***
semi_skilled  0.57587    0.11264   5.112 3.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000269)

Number of Fisher Scoring iterations: 4
```

These results are now the same as in the published model

I now save the (latest) data frame as a file "ycs9sw1_v2.rda".

In [25]:

```
save(mydata4.df,file="C:/Users/Vernon/OneDrive - University of Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.rda")
```

List the objects in my workspace.

In [26]:

```
ls()
```

Out[26]:

"model1" "model2" "mydata.df" "mydata2.df" "mydata3.df" "mydata4.df" "mydesign1" "mydesign2"

Now I am going to remove "rm" these objects.

In [27]:

```
rm ("mydata.df", "mydata2.df", "mydata3.df", "model1", "mydesign1")
ls()
```

Out[27]:

```
"model2"   "mydata4.df"   "mydesign2"
```

Duplicating the Connelly (2006) Model Results in SPSS

A close look at the results of the model in *R* indicate that whilst the values of the parameter estimates "estimates" are the same as *B* in Table 5 (p.20) the standard errors are not the same.

My intuition is that the original analysis was undertaken in **SPSS**.

This is an unforeseen obstacle.

My desire is to investigate this a little further.

Unfortunately, at the current time there is not a SPSS kernel available within the Jupyter notebook.

However, I have a cunning plan.

First, I will write out the dataset in SPSS format.

write.foreign doesn't generate native SPSS datafiles (.sav) but it does generate is the data in a comma delimited format (a .txt file) and a basic syntax file for reading that data into SPSS (a .sps file).

Using the following general syntax

write.foreign(as.data.frame(mydata), "c:/mydata.txt", "c:/mydata.sps", package="SPSS")

I plan to estimate the logit model in SPSS (with the data weighted).

In [28]:

```
write.foreign(as.data.frame(mydata4.df), "C:/Users/Vernon/OneDrive - Univer
sity of Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.txt",
"C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.sps", package="SPSS")
```

```
Warning message:
In writeForeignSPSS(df = structure(list(serial = c(200001L, 200004L, : some
variable names were abbreviated
```

Here we leave the Jupyter notebook

We have had to break the workflow because SPSS cannot currently be run in the language agnostic environment of the Jupyter notebook.

To assist with transparency this link shows the model being estimated in SPSS
(IBM SPSS Version 22 Release 22.0.0.1 64 bit)

(IBM SPSS Version 22 Release 22.0.0.1 64-bit)

https://youtu.be/12YXww67m9s

I am grateful to Dr Roxanne Connelly, University of Warwick, UK (http://www2.warwick.ac.uk/fac/soc/sociology/staff/connelly/) for this suggestion.

I use the following on-line software to record the SPSS job https://www.apowersoft.com/free-online-screen-recorder .

Here is the SPSS syntax that was generated by the *write.foreign* command.

```
DATA LIST FILE= "C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.txt" free (",") / serial weight sex s1a_c a58 s1eth
s1acqe pseg prof_man o_non_mn sklld_mn sm_sklld pseg5 pseg6 pseg7 s15a_c girls ethnic1 white
black indian pakistan bangldsh chinese other .

VARIABLE LABELS serial "serial" weight "weight" sex "sex" s1a_c "s1a_c" a58 "a58" s1eth "s1eth"
s1acqe "s1acqe" pseg "pseg" prof_man "prof_man" o_non_mn "o_non_man" sklld_mn "skilled_man"
sm_sklld "semi_skilled" pseg5 "pseg5" pseg6 "pseg6" pseg7 "pseg7" s15a_c "s15a_c" girls "girls"
ethnic1 "ethnic1" white "white" black "black" indian "indian" pakistan "pakistani" bangldsh
"bangladeshi" chinese "chinese" other "other" .

VALUE LABELS / sex 1 "not answered (9)" 2 "item not applicable" 3 "male" 4 "female" / a58 1 "not
answered (99)" 2 "don't know (98)" 3 "item not applicable" 4 "white" 5 "carib." 6 "afro." 7 "other black"
8 "indian" 9 "pakistani" 10 "bangladeshi" 11 "chin!ese" 12 "other asian" 13 "any other" 14 "mixed
ethnic origin" 15 "ref!used" / s1eth 1 "not answered (9)" 2 "schedule not obtained" 3 "schedule not
applicable" 4 "item not applicable" 5 "white" 6 "black groups" 7 "asian groups" 8 "mixed/!other groups"
9 "refused/ns" / s1acqe 1 "not answered (9)" 2 "schedule not obtained" 3 "schedule not applicable" 4
"item not applicable" 5 "5+ gcses at a-c" 6 "1-4 gcses at a-c" 7 "5+ gcses at d-g" 8 "1-4 gcses at d-g"
9 "none reported" . EXECUTE.
```

Here is the SPSS syntax that weights the data and then estimates the logit model.

```
WEIGHT BY weight.

LOGISTIC REGRESSION VARIABLES s15a_c /METHOD=ENTER girls chinese indian white
bangldsh pakistan prof_man o_non_mn sklld_mn sm_sklld /CRITERIA=PIN(.05) POUT(.10)
ITERATE(20) CUT(.5).
```

These results are now the same as in the published model

Complex Samples Logistic Regression

Complex Samples Logistic Regression.

As a final check I undertake the analysis in SPSS using the Comples Samples approach.

The code required for the complex sample analysis plans is

```
CSPLAN ANALYSIS
/PLAN FILE='C:\Users\Vernon\OneDrive - University of
Edinburgh\Documents\ycs_9_2017\logit.csaplan'
/PLANVARS ANALYSISWEIGHT=weight
/SRSESTIMATOR TYPE=WOR
/PRINT PLAN
/DESIGN
/ESTIMATOR TYPE=WR.
```

The code required for the complex sample logistic regression model is

```
CSLOGISTIC s15a_c(LOW) WITH girls chinese indian white pakistan bangldsh prof_man o_non_mn
skIld_mn sm_skIld
/PLAN FILE='C:\Users\Vernon\OneDrive - University of '+
'Edinburgh\Documents\ycs_9_2017\logit.csaplan'
/MODEL girls chinese indian white bangldsh pakistan prof_man o_non_mn skIld_mn sm_skIld
/INTERCEPT INCLUDE=YES SHOW=YES
/STATISTICS PARAMETER SE
/TEST TYPE=F PADJUST=LSD
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE] LCONVERGE=[0]
CHKSEP=20 CILEVEL=95
/PRINT SUMMARY CLASSTABLE VARIABLEINFO SAMPLEINFO.
```

Here is a screen shot of the SPSS output.

There results are the same as the results in *R*

It is worth noting that there are (at least) two ways of estimating a logit model in SPSS in the presence of survey weights.

The __Complex Samples__ approach returns the same results as __svy__ in _R_.

By contrast weighting the dataset first, and then estimating a standard logistic regression model leads to different standard errors.

Detective work was required to arrive at this conclusion.

This passage of work underlines the requirement to clearly state the software used (including versions, libraries and dependencies) and as much detail as possible relating to the technique used.

Duplicating the Connolly (2006) Model Results in Stata

In this section I duplicate the results produced in Connolly 2006 using Stata.

We have had to move outside of the workflow in Jupyter (to move to SPSS).

Just in case we should make certain that we have the correct dataset in the frame.

In [4]:

```
load ("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.rda")
ls()
```

Out[4]:

"mydata4.df"

I will now export data frame to Stata format using the *foreign* library.

In [7]:

```
library(foreign)
write.dta(mydata4.df, "C:/Users/Vernon/OneDrive - University of Edinburgh/D
ocuments/ycs_9_2017/ycs9sw1_v2.dta")
```

You MUST have Stata on your machine!

This section uses *ipystata* to run Stata via Jupyter magic see http://dev-ii-seminar.readthedocs.io/en/latest/notebooks/Stata_in_jupyter.html .

You can install IPyStata 0.3.0 using the following syntax

pip install ipystata

at your command line prompt i.e. c:\Users\Vernon .

This facility is provided here

https://github.com/TiesdeKok/ipystata

The author is Ties de Kok
e-mail: t.c.j.dekok@tilburguniversity.edu
Twitter: @TiesdeKok

You MUST have Stata on your machine!

This cell below imports ipystata so that we can run Stata within this notebook.

You MUST have Stata on your machine!

You MUST CHANGE the Jupyter kernel to PYTHON

Use the Kernel menu above.

Python is native to Jupyter so you will have this kernel.

In [1]:

```
import ipystata
```

If you have an error that looks a bit like this

Error in parse(text = x, srcfile = src): 1:8: unexpected symbol
1: import ipystata

Then you may probably have changed kernel

You MUST CHANGE the Jupyter kernel to PYTHON using the drop down menu Kernel above.

We are now working in Python using Stata via magic cells!

In [2]:

```
%%stata -o mydata4.df
codebook, compact
```

```
Variable        Obs Unique      Mean     Min      Max  Label
------------------------------------------------------------
serial        12789  12789  212370.3  200001   231392  se...
weight        12789    178  .9822923  .60253  2.51757  we...
sex           12789      2  3.532411       3        4  sex
s1a_c         12789     14  5.526624       0       13  s1a_c
a58           12789      8  4.265619       4       11  a58
s1eth         12789      3  5.115255       5        7  s1eth
s1acqe        12789      5  5.651576       5        9  s1...
pseg          12789      1         0       0        0  pseg
prof_man      12789      2  .2561576       0        1  pr...
o_non_man     12789      2  .2298851       0        1  o_...
skilled_man   12789      2  .3528814       0        1  sk...
semi_skilled  12789      2  .1215107       0        1  se...
pseg5         12789      2  .0395653       0        1  pseg5
pseg6         12789      1         0       0        0  pseg6
pseg7         12789      1         0       0        0  pseg7
s15a_c        12789      2  .6023927       0        1  s1...
girls         12789      2  .5324107       0        1  girls
ethnic1       12789      6  1.151536       1        6  et...
white         12789      2  .9353351       0        1  white
black         12789      2  .0140746       0        1  black
indian        12789      2  .0288529       0        1  in...
pakistani     12789      2  .0124326       0        1  pa...
bangladeshi   12789      2   .004066       0        1  ba...
chinese       12789      2  .0052389       0        1  ch...
other         12789      1         0       0        0  other
```

```
--------------------------------------------------------------
> ta", replace
```

In [6]:

```
%%stata -o mydata4.df
svyset [pweight=weight]
* you may need to set the line size to stop the table going wonky *
set linesize 100
svy:logit s15a_c girls chinese indian white pakistani bangladeshi prof_man
o_non_man skilled_man semi_skilled
```

```
      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
> emi_skilled
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =         1                Number of obs      =
12789
Number of PSUs     =     12789                Population size    =
12562.536
                                              Design df          =     12
                                              F(  10,  12779)    =
113.58
                                              Prob > F           =    0.0

------------------------------------------------------------------------------
             |             Linearized
      s15a_c |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interv
l]
-------------+----------------------------------------------------------------
       girls |   .4045638   .0392609    10.30   0.000     .3276067     .481
21
     chinese |   2.002307   .3773355     5.31   0.000     1.262673    2.741
41
      indian |   1.065842   .2082879     5.12   0.000      .657567    1.474
18
       white |   .6431364   .1711832     3.76   0.000     .3075918     .978
81
   pakistani |   .5313644   .2450316     2.17   0.030     .0510659    1.011
63
 bangladeshi |   .7661585   .3448563     2.22   0.026     .0901887
1.442128
    prof_man |   2.192092   .1086303    20.18   0.000      1.97916    2.405
23
   o_non_man |    1.77251   .1079307    16.42   0.000      1.56095     1.98
07
 skilled_man |   .9321662   .1041115     8.95   0.000      .728092     1.13
24
semi_skilled |   .5758727   .1126428     5.11   0.000      .355076    .7966
93
       _cons |  -2.208288   .1980157   -11.15   0.000    -2.596429   -1.820
48
------------------------------------------------------------------------------
```

Here are the result from *R*

Survey design: svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

variable	Estimate	Std. Error	t value	Pr
(Intercept)	-2.20829	0.19802	-11.152	< 2e-16
girls	0.40456	0.03926	10.305	< 2e-16
chinese	2.00231	0.37734	5.306	1.14e-07
indian	1.06584	0.20829	5.117	3.15e-07
white	0.64314	0.17118	3.757	0.000173
bangladeshi	0.76616	0.34486	2.222	0.026323
pakistani	0.53136	0.24503	2.169	0.030135
prof_man	2.19209	0.10863	20.179	< 2e-16
o_non_man	1.77251	0.10793	16.423	< 2e-16
skilled_man	0.93217	0.10411	8.954	< 2e-16
semi_skilled	0.57587	0.11264	5.112	3.23e-07

Both their coefficients and the standard errors are the same in __ _R_ __ and in __Stata__

You MUST CHANGE the Jupyter kernel to _R_

Use the Kernel menu above.

The *R* kernel is required as we are moving back to *R* .

In this section I plan to export the data frame "mydata4.df" which is in the file "ycs9sw1_v2" into an Excel format file ".xlsx".

This required the package 'xlsx' to be installed in *R* .

> *install.packages('xlsx')*

When I first tried this there was an error because a more up to date version of Java is required.

In [1]:

```
# if the work recommences with this section then the following libraries mi
ght be required.
library(foreign)
library(survey)
library(car)
library(car)
library(dplyr)
library(weights)
library(dummies)
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package:
grid
Loading required package: Matrix
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns

Warning message:
: package 'xlsx' was built under R version 3.2.5Loading required package: r
Java
Warning message:
: package 'rJava' was built under R version 3.2.5Loading required package:
xlsxjars
Warning message:
: package 'xlsxjars' was built under R version 3.2.5
```

In [ ]:

```
# this library is required.
library(xlsx)
```

In [2]:

```
load ("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.rda")
ls()
```

Out[2]:

"mydata4.df"

In [3]:

```
write.xlsx(mydata4.df, "C:/Users/Vernon/OneDrive - University of Edinburgh/
Documents/ycs_9_2017/ycs9sw1_v2.xlsx")
```

A file called "ycs9sw1_v2.xlsx" has now been written from within *R*.

Duplicating the Connolly (2006) Model Results in Python

In this section I will attempt to duplicate the logit model Table 5 p.20 Connolly (2006) in Python.

You MUST CHANGE the Jupyter kernel to PYTHON

Use the Kernel menu above.

Python is native to Jupyter so you will have this kernel.

First we have to "import" a package called "pandas".

Pandas is a software library written for the Python programming language for data manipulation and analysis.

In [1]:

```
import pandas as pd
```

If you have an error that looks a bit like this

Error in parse(text = x, srcfile = src): :1:8: unexpected symbol 1: import pandas

Then you may probably have changed kernel

You MUST CHANGE the Jupyter kernel to PYTHON using the drop down menu Kernel above

Using "read_excel" which is part of pandas which I have already loaded into "pd", I now construct the data frame "df" reading in the data from the Excel (xlsx) file.

In [2]:

```
df = pd.read_excel("C:/Users/Vernon/OneDrive - University of Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.xlsx")
df.head()
```

Out[2]:

	serial	weight	sex	s1a_c	a58	s1eth	s1acqe	pseg	prof_man	o_non_man	...	s15
1	200001	0.87518	female	9	white	white	5+ gcses at a*-c	0	1	0	...	1
2	200004	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	...	1
3	200005	0.97615	male	9	indian	asian groups	5+ gcses at a*-c	0	0	0	...	1
4	200006	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	...	1
8	200014	0.95928	female	2	white	white	1-4 gcses at a*-c	0	0	0	...	0

5 rows × 25 columns

Python is more general purpose and not primarily orientated towards social science data analysis. Therefore some things are a little more fiddly.

For example before estimating the logistic regression models we must set a constant for all case (int=1).

In [3]:

```
df['Int']=1
```

Examining the data in the data frame "df".

In [20]:

```
df.head()
df.head()
```

Out[20]:

	serial	weight	sex	s1a_c	a58	s1eth	s1acqe	pseg	prof_man	o_non_man	...	s15
1	200001	0.87518	female	9	white	white	5+ gcses at a*-c	0	1	0	...	1
2	200004	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	...	1
3	200005	0.97615	male	9	indian	asian groups	5+ gcses at a*-c	0	0	0	...	1
4	200006	0.97615	male	9	white	white	5+ gcses at a*-c	0	0	1	...	1
8	200014	0.95928	female	2	white	white	1-4 gcses at a*-c	0	0	0	...	0

5 rows × 25 columns

In [21]:

```
df.describe()
```

Out[21]:

	serial	weight	s1a_c	pseg	prof_man	o_non_man	skillec
count	12789.000000	12789.000000	12789.000000	12789	12789.000000	12789.000000	12789
mean	212370.330988	0.982292	5.526624	0	0.256158	0.229885	0.3528
std	7442.695401	0.350797	3.671873	0	0.436527	0.420775	0.4778
min	200001.000000	0.602530	0.000000	0	0.000000	0.000000	0.0000
25%	206648.000000	0.762780	2.000000	0	0.000000	0.000000	0.0000
50%	211922.000000	0.875030	6.000000	0	0.000000	0.000000	0.0000
75%	217230.000000	1.030390	9.000000	0	1.000000	0.000000	1.0000
max	231392.000000	2.517570	13.000000	0	1.000000	1.000000	1.0000

8 rows × 21 columns

In [3]:

```
import statsmodels.api as sm
```

In [6]:

```
list(df)
```

Out[6]:

```
['serial',
 'weight',
 'sex',
 's1a_c',
 'a58',
 's1eth',
 's1acqe',
 'pseg',
 'prof_man',
 'o_non_man',
 'skilled_man',
 'semi_skilled',
 'pseg5',
 'pseg6',
 'pseg7',
 's15a_c',
 'girls',
 'ethnic1',
 'white',
 'black',
 'indian',
 'pakistani',
 'bangladeshi',
 'chinese',
 'other',
 'Int']
```

In [7]:

```
independentVar = ['girls', 'chinese', 'indian', 'white', 'bangladeshi', 'pa
kistani', 'prof_man','o_non_man','skilled_man','semi_skilled', 'Int']
logReg = sm.Logit(df['s15a_c'] , df[independentVar])
answer = logReg.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.625258
         Iterations 5
```

In [8]:

```
answer.summary()
```

Out[8]:

Logit Regression Results

Dep. Variable:	s15a_c	**No. Observations:**	12789
Model:	Logit	**Df Residuals:**	12778
Method:	MLE	**Df Model:**	10
Date:	Thu, 22 Jun 2017	**Pseudo R-squ.:**	0.06960
Time:	16:57:45	**Log-Likelihood:**	-7996.4
converged:	True	**LL-Null:**	-8594.6
		LLR p-value:	8.895e-251

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
girls	0.3239	0.038	8.519	0.000	0.249 0.398
chinese	1.8696	0.347	5.393	0.000	1.190 2.549
indian	1.0400	0.195	5.331	0.000	0.658 1.422
white	0.6486	0.158	4.095	0.000	0.338 0.959
bangladeshi	0.7172	0.330	2.173	0.030	0.070 1.364
pakistani	0.4700	0.228	2.059	0.040	0.023 0.917
prof_man	2.0805	0.106	19.629	0.000	1.873 2.288
o_non_man	1.6869	0.105	16.007	0.000	1.480 1.893
skilled_man	0.8715	0.102	8.564	0.000	0.672 1.071
semi_skilled	0.5475	0.110	4.972	0.000	0.332 0.763
Int	-1.6659	0.186	-8.962	0.000	-2.030 -1.302

In [8]:

```
from patsy import dmatrices
```

In [9]:

```
y, x = dmatrices('s15a_c ~ 1 + girls + chinese + indian + white + bangladeshi + pakistani + prof_man + o_non_man + skilled_man + semi_skilled' , df)
```

In [10]:

```
sm.GLM(endog=y, exog=x, family=sm.families.Binomial(), data_weights=df['weight']).fit().summary()
```

Out[10]:

Generalized Linear Model Regression Results

Dep. Variable:	s15a_c	**No. Observations:**	12789
Model:	GLM	**Df Residuals:**	12778
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0
Method:	IRLS	**Log-Likelihood:**	-7996.4
Date:	Mon, 26 Jun 2017	**Deviance:**	15993.
Time:	12:50:00	**Pearson chi2:**	1.28e+04
No. Iterations:	6		

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	-1.6659	0.186	-8.962	0.000	-2.030 -1.302
girls	0.3239	0.038	8.519	0.000	0.249 0.398
chinese	1.8696	0.347	5.393	0.000	1.190 2.549

indian	1.0400	0.195	5.331	0.000	0.658 1.422
white	0.6486	0.158	4.095	0.000	0.338 0.959
bangladeshi	0.7172	0.330	2.173	0.030	0.070 1.364
pakistani	0.4700	0.228	2.059	0.040	0.023 0.917
prof_man	2.0805	0.106	19.629	0.000	1.873 2.288
o_non_man	1.6869	0.105	16.007	0.000	1.480 1.893
skilled_man	0.8715	0.102	8.564	0.000	0.672 1.071
semi_skilled	0.5475	0.110	4.972	0.000	0.332 0.763

Beware!

These results do not appear to have been weighted.

In [14]:

```
# Here is another attempt...

logmodel=sm.GLM(endog=y, exog=x,
family=sm.families.Binomial(sm.families.links.logit)).fit()
#sm.GLM(, family=sm.families.Binomial(),
data_weights=df['weight']).fit().summary()
logmodel.summary()
```

Out[14]:

Generalized Linear Model Regression Results

Dep. Variable:	s15a_c	**No. Observations:**	12789
Model:	GLM	**Df Residuals:**	12778
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0
Method:	IRLS	**Log-Likelihood:**	-7996.4
Date:	Mon, 26 Jun 2017	**Deviance:**	15993.
Time:	12:54:07	**Pearson chi2:**	1.28e+04
No. Iterations:	6		

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	-1.6659	0.186	-8.962	0.000	-2.030 -1.302
girls	0.3239	0.038	8.519	0.000	0.249 0.398
chinese	1.8696	0.347	5.393	0.000	1.190 2.549
indian	1.0400	0.195	5.331	0.000	0.658 1.422

white	0.6486	0.158	4.095	0.000	0.338 0.959
bangladeshi	0.7172	0.330	2.173	0.030	0.070 1.364
pakistani	0.4700	0.228	2.059	0.040	0.023 0.917
prof_man	2.0805	0.106	19.629	0.000	1.873 2.288
o_non_man	1.6869	0.105	16.007	0.000	1.480 1.893
skilled_man	0.8715	0.102	8.564	0.000	0.672 1.071
semi_skilled	0.5475	0.110	4.972	0.000	0.332 0.763

Beware!

These results do not appear to have been weighted.

In [17]:

```
#Here is a third attempt ...

weight =df['weight']
logmodel2=sm.GLM(endog=y, exog=x, sample_weight =weight,
family=sm.families.Binomial(sm.families.links.logit)).fit()

#sm.GLM(, family=sm.families.Binomial(),
data_weights=df['weight']).fit().summary()
logmodel2.summary()
```

Out[17]:

Generalized Linear Model Regression Results

Dep. Variable:	s15a_c	**No. Observations:**	12789
Model:	GLM	**Df Residuals:**	12778
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0
Method:	IRLS	**Log-Likelihood:**	-7996.4
Date:	Mon, 26 Jun 2017	**Deviance:**	15993.
Time:	13:00:21	**Pearson chi2:**	1.28e+04
No. Iterations:	6		

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	-1.6659	0.186	-8.962	0.000	-2.030 -1.302
girls	0.3239	0.038	8.519	0.000	0.249 0.398
chinese	1.8696	0.347	5.393	0.000	1.190 2.549
indian	1.0400	0.195	5.331	0.000	0.658 1.422

white	0.6486	0.158	4.095	0.000	0.338 0.959
bangladeshi	0.7172	0.330	2.173	0.030	0.070 1.364
pakistani	0.4700	0.228	2.059	0.040	0.023 0.917
prof_man	2.0805	0.106	19.629	0.000	1.873 2.288
o_non_man	1.6869	0.105	16.007	0.000	1.480 1.893
skilled_man	0.8715	0.102	8.564	0.000	0.672 1.071
semi_skilled	0.5475	0.110	4.972	0.000	0.332 0.763

Beware!

These results do not appear to have been weighted.

In order to investigate this further I will return to *R* .

Change the kernel back to *R* .

In [1]:

```
library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)
library(xlsx)
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package:
grid
Loading required package: Matrix
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':
    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns

Warning message:
: package 'xlsx' was built under R version 3.2.5Loading required package: r
Java
Warning message:
: package 'rJava' was built under R version 3.2.5Loading required package:
xlsxjars
Warning message:
: package 'xlsxjars' was built under R version 3.2.5
```

In [1]:

```
load("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.rda")
ls()
```

Out[1]:

"mydata4.df"

In [2]:

```
summary(mydata4.df)
```

Out[2]:

```
     serial            weight                                 sex            s1a_c

 Min.   :200001   Min.   :0.6025   not answered (9)    :    0   Min.   :
0.000
 1st Qu.:206648   1st Qu.:0.7628   item not applicable:    0   1st Qu.: 2.00
0
 Median :211922   Median :0.8750   male                :5980   Median : 6.00

 Mean   :212370   Mean   :0.9823   female              :6809   Mean   : 5.52

 3rd Qu.:217230   3rd Qu.:1.0304                                3rd Qu.: 9.00
 Max.   :231392   Max.   :2.5176                                Max.   :13.00
```

```
         a58                                  s1eth                     s1acqe
white      :11962   white                   :11962   5+ gcses at a*-c :7747

indian     :  369   asian groups            :  647   1-4 gcses at a*-c:3071

pakistani  :  159   black groups            :  180   5+ gcses at d-g  :1190

carib.     :   73   not answered (9)        :    0   none reported    : 539

chin!ese   :   67   schedule not obtained   :    0   1-4 gcses at d-g : 242
other black:   58   schedule not applicable:    0   not answered (9) :   0

(Other)    :  101   (Other)                 :    0   (Other)          :   0
     pseg      prof_man        o_non_man       skilled_man
Min.   :0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0   Mean   :0.2562   Mean   :0.2299   Mean   :0.3529
3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

 semi_skilled        pseg5              pseg6        pseg7        s15a_c
Min.   :0.0000   Min.   :0.00000   Min.   :0   Min.   :0   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0   1st Qu.:0   1st Qu.:0.0000

Median :0.0000   Median :0.00000   Median :0   Median :0   Median :1.0000

Mean   :0.1215   Mean   :0.03957   Mean   :0   Mean   :0   Mean   :0.6024
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1.0000

Max.   :1.0000   Max.   :1.00000   Max.   :0   Max.   :0   Max.   :1.0000

    girls           ethnic1          white             black
Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.00000
Median :1.0000   Median :1.000   Median :1.0000   Median :0.00000
Mean   :0.5324   Mean   :1.152   Mean   :0.9353   Mean   :0.01407
3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :6.000   Max.   :1.0000   Max.   :1.00000

    indian            pakistani          bangladeshi          chinese
Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
Median :0.00000   Median :0.00000   Median :0.000000   Median :0.000000
Mean   :0.02885   Mean   :0.01243   Mean   :0.004066   Mean   :0.005239
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
Max.   :1.00000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000

    other
Min.   :0
1st Qu.:0
Median :0
Mean   :0
3rd Qu.:0
Max.   :0
```

In order to check the results produced using Python I will re-estimate the model in *R* but this time ignoring the sample weights.

In [3]:

```
modelnw<-glm (s15a_c ~ girls + chinese + indian +
                 white + bangladeshi + pakistani +
                 pakistani + prof_man + o_non_man +
                 skilled_man + semi_skilled, data = mydata4.df, family = "bi
nomial")
```

In [4]:

```
summary(modelnw)
```

Out[4]:

```
Call:
glm(formula = s15a_c ~ girls + chinese + indian + white + bangladeshi +
    pakistani + pakistani + prof_man + o_non_man + skilled_man +
    semi_skilled, family = "binomial", data = mydata4.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1823  -1.1162   0.6678   0.9093   1.7744

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.66590    0.18588  -8.962  < 2e-16 ***
girls         0.32387    0.03802   8.519  < 2e-16 ***
chinese       1.86961    0.34667   5.393 6.93e-08 ***
indian        1.04002    0.19507   5.331 9.75e-08 ***
white         0.64860    0.15840   4.095 4.23e-05 ***
bangladeshi   0.71721    0.33011   2.173   0.0298 *
pakistani     0.46998    0.22830   2.059   0.0395 *
prof_man      2.08045    0.10599  19.629  < 2e-16 ***
o_non_man     1.68694    0.10539  16.007  < 2e-16 ***
skilled_man   0.87151    0.10177   8.564  < 2e-16 ***
semi_skilled  0.54747    0.11011   4.972 6.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17189  on 12788  degrees of freedom
Residual deviance: 15993  on 12778  degrees of freedom
AIC: 16015

Number of Fisher Scoring iterations: 4
```

These un-weighted results are the same as the *Python* results. The weighting in not working in in *Python*.

Further investigation of how to incorporate survey weights into a logistic regression model using *Python* is required.

Replicating the Connolly (2006) Model Results with Quasi-Variance

In this section, following the *duplication* of the logistic rgression results in Table 3 (p.20) of Connolley (2006) I now undertake a *replication* activity.

In brief, I have concerns about the parameterisation of the ethnicity measure in the logistic regression model.

The reference category is 'Black' pupils.

This is a small category (*n*=180).

My suspicion is that this is a sub-optimal reference category.

I will investigate the relationship between the categories of ethnicity by estimating quasi-variance based comparison intervals.

An extensive and reproducible introduction is provided by Gayle and Lambert (2007).

The use of quasi-variance based comparison intervals allows a more subtle investigation of the differences between ethnic groups.

The procedure that will be used is described in Firth and De Menezes (2004) an implemented in the *R* library 'qvcalc'.

To run this procedure you must have __qvcalc__ installed in _R_.

code required in *R*
install.packages('qvcalc')

Warning.

Within this Jupyter notebook there has been a lot of non-routine work. For example I have 'swivel-chaired' between data analytical software packages and changed kernels.

It may from time to time be necessary to re-start the notebook depending on how stable your computing environment is.

In this section I re-start a *R* session.

In [1]:

```
library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)
library(MASS)
library(qvcalc)
```

```
Warning message:
: package 'survey' was built under R version 3.2.5Loading required package:
grid
Loading required package: Matrix
Loading required package: survival
Warning message:
: package 'survival' was built under R version 3.2.5
Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart

Warning message:
: package 'car' was built under R version 3.2.5Warning message:
: package 'dplyr' was built under R version 3.2.5
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
lattice
Loading required package: Formula
Warning message:
: package 'Formula' was built under R version 3.2.5Loading required package
: ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

```
Warning message:
: package 'dummies' was built under R version 3.2.5dummies-1.5.6 provided b
y Decision Patterns


Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select

Warning message:
: package 'qvcalc' was built under R version 3.2.5
```

Various WARNINGS will appear. Don't panic.

I re-load the *R* file "ycs9sw1_v2.rda".

```
In [2]:
```

```
load("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v2.rda")
ls()
```

```
Out[2]:
```

"mydata4.df"

The data frame is "mydata4.df".

Please double check that an earlier version has not bee loaded!

```
In [3]:
```

```
summary(mydata4.df)
```

```
Out[3]:
```

```
     serial           weight                          sex              s1a_c

 Min.   :200001   Min.   :0.6025   not answered (9)    :    0   Min.   :
0.000
 1st Qu.:206648   1st Qu.:0.7628   item not applicable:    0   1st Qu.: 2.00
0
 Median :211922   Median :0.8750   male               :5980   Median : 6.00

 Mean   :212370   Mean   :0.9823   female             :6809   Mean   : 5.52

 3rd Qu.:217230   3rd Qu.:1.0304                               3rd Qu.: 9.00
 Max.   :231392   Max.   :2.5176                               Max.   :13.00



          a58                           s1eth                         s1acqe
 white       :11962   white                     :11962   5+ gcses at a*-c :7747

 indian      :  369   asian groups              :  647   1-4 gcses at a*-c:3071

 pakistani   :  159   black groups              :  180   5+ gcses at d-g  :1190

 carib.      :   73   not answered (9)          :    0   none reported    : 539

 chin!ese    :   67   schedule not obtained     :    0   1-4 gcses at d-g : 242
 other black:   58   schedule not applicable:    0   not answered (9) :    0

 (Other)     :  101   (Other)                   :    0   (Other)          :    0
      pseg        prof_man        o_non_man        skilled_man
 Min.   :0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0   Mean   :0.2562   Mean   :0.2299   Mean   :0.3529
 3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
  semi_skilled        pseg5              pseg6       pseg7        s15a_c      
 Min.   :0.0000   Min.   :0.00000   Min.   :0   Min.   :0   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0   1st Qu.:0   1st Qu.:0.0000  

 Median :0.0000   Median :0.00000   Median :0   Median :0   Median :1.0000  

 Mean   :0.1215   Mean   :0.03957   Mean   :0   Mean   :0   Mean   :0.6024  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1.0000  

 Max.   :1.0000   Max.   :1.00000   Max.   :0   Max.   :0   Max.   :1.0000  

     girls           ethnic1          white            black        
 Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.00000  
 Median :1.0000   Median :1.000   Median :1.0000   Median :0.00000  
 Mean   :0.5324   Mean   :1.152   Mean   :0.9353   Mean   :0.01407  
 3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :6.000   Max.   :1.0000   Max.   :1.00000  

     indian            pakistani         bangladeshi          chinese        
 Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000  
 Median :0.00000   Median :0.00000   Median :0.000000   Median :0.000000  
 Mean   :0.02885   Mean   :0.01243   Mean   :0.004066   Mean   :0.005239  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000  

     other  
 Min.   :0  
 1st Qu.:0  
 Median :0  
 Mean   :0  
 3rd Qu.:0  
 Max.   :0  
```

I now re-estimate the logit model that "duplicated" the results in Table 3 (p.20) Connolly (2016).

In [4]:

```
mydesign2 <- svydesign(id = ~serial,data = mydata4.df, weight = ~weight)
```

In [5]:

```
model2<-svyglm (s15a_c ~ girls + chinese + indian +
                  white + bangladeshi + pakistani +
                  pakistani + prof_man + o_non_man +
                  skilled_man + semi_skilled, design=mydesign2, data = mydat
4.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

In [6]:

```
summary(model2)
```

Out[6]:

Out[6]:

```
Call:
svyglm(formula = s15a_c ~ girls + chinese + indian + white + 
    bangladeshi + pakistani + pakistani + prof_man + o_non_man + 
    skilled_man + semi_skilled, design = mydesign2, data = mydata4.df, 
    family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.20829    0.19802 -11.152  < 2e-16 ***
girls         0.40456    0.03926  10.305  < 2e-16 ***
chinese       2.00231    0.37734   5.306 1.14e-07 ***
indian        1.06584    0.20829   5.117 3.15e-07 ***
white         0.64314    0.17118   3.757 0.000173 ***
bangladeshi   0.76616    0.34486   2.222 0.026323 *  
pakistani     0.53136    0.24503   2.169 0.030135 *  
prof_man      2.19209    0.10863  20.179  < 2e-16 ***
o_non_man     1.77251    0.10793  16.423  < 2e-16 ***
skilled_man   0.93217    0.10411   8.954  < 2e-16 ***
semi_skilled  0.57587    0.11264   5.112 3.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000269)

Number of Fisher Scoring iterations: 4
```

Here is a reminder of the variables that are in the data frame "mydata4.df".

In [7]:

```
str(mydata4.df)
```

```
'data.frame':	12789 obs. of  25 variables:
 $ serial      : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight      : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex         : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3 
4 4 ...
 $ s1a_c       : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58         : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4 
4 4 4 ...
 $ s1eth       : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5 
5 5 ...
 $ s1acqe      : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6 
5 6 ...
 $ pseg        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c       : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls        : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1      : num  1 1 3 1 1 1 1 1 1 1 ...
 $ white        : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian       : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other        : num  0 0 0 0 0 0 0 0 0 0 ...
```

In order to use the QV procedure I have to estimate the model with a multiple-categorie measure of "ethnicity".

The variable "ethnicity1" has already been created.

I check that "ethnicity1" is a factor.

In [8]:

```
levels(mydata4.df$ethnic1)
```

Out[8]:

```
NULL
```

The variable "ethnic1" is not a factor so I am going to declare as a factor.

In [9]:

```
mydata4.df$ethnic1  <- factor(mydata4.df$ethnic1 )
```

In [10]:

```
levels(mydata4.df$ethnic1)
```

Out[10]:

```
"1"  "2"  "3"  "4"  "5"  "6"
```

In [11]:

```
is.factor(mydata4.df$ethnic1)
```

Out[11]:

TRUE

The variable "ethnic1" is now a factor.

Here I remind myself of the variables in the data frame "mydata4.df" and check again that "ethnic1" is a factor in the dataset.

In [13]:

```
str(mydata4.df)
```

```
'data.frame':  12789 obs. of  25 variables:
```

```
'data.frame':	12709 obs. of  25 variables:
 $ serial      : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight      : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex         : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c       : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58         : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth       : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe      : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls       : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1     : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white       : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
```

In [14]:

```
ls()
```

Out[14]:

"model2" "mydata4.df" "mydesign2"

In [15]:

```
model3<-svyglm (s15a_c ~ factor(ethnic1) + girls + prof_man + o_non_man +
                 skilled_man + semi_skilled, design=mydesign2, data = mydata
4.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

In [16]:

```
summary(model3)
```

Out[16]:

```
Call:
svyglm(formula = s15a_c ~ factor(ethnic1) + girls + prof_man + 
    o_non_man + skilled_man + semi_skilled, design = mydesign2, 
    data = mydata4.df, family = "binomial")

Survey design:
```

```
Survey design:
svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.56515    0.10196 -15.351  < 2e-16 ***
factor(ethnic1)2 -0.64314    0.17118  -3.757 0.000173 ***
factor(ethnic1)3  0.42271    0.12187   3.469 0.000525 ***
factor(ethnic1)4 -0.11177    0.17735  -0.630 0.528544    
factor(ethnic1)5  0.12302    0.30052   0.409 0.682278    
factor(ethnic1)6  1.35917    0.33721   4.031 5.59e-05 ***
girls             0.40456    0.03926  10.305  < 2e-16 ***
prof_man          2.19209    0.10863  20.179  < 2e-16 ***
o_non_man         1.77251    0.10793  16.423  < 2e-16 ***
skilled_man       0.93217    0.10411   8.954  < 2e-16 ***
semi_skilled      0.57587    0.11264   5.112 3.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000269)

Number of Fisher Scoring iterations: 4
```

BEWARE

The variable "ethnic1" is coded to match the ethnicity measure in Table 1 (p.7) Connolly (2006). **However**, the order of the dummy variables included in the logistic regression model in Table 3 (p.20) Connelly (2006) do not match.

This could not have easily been foreseen.

In the spirit of showing **all** of the workflow I have preseved this snippet of data wrangling.

A re-coded version of "ethnic1" is required.

Here is the original variable.

In [18]:

```
table(mydata4.df$ethnic1)
```

Out[18]:

```
    1     2     3     4     5     6 
11962   180   369   159    52    67 
```

The new variable will be called "ethnic2".

The reference category should be 'black' pupils (i.e. carib; afro.; other black).

Categories should be

1. Black
2. Chinese
3. Indian

4. White
5. Bangladeshi
6. Pakistani
7. Others (but this category has been omitted from the analysis)

In [19]:

```
# create the new field,

mydata4.df$ethnic2 <- 7

# everyone is placed into category 7.

# recode the new field with values from the old field.
mydata4.df$ethnic2[mydata4.df$a58=="white"] <-4
mydata4.df$ethnic2[mydata4.df$a58=="carib."] <-1
mydata4.df$ethnic2[mydata4.df$a58=="afro."] <-1
mydata4.df$ethnic2[mydata4.df$a58=="other black"] <-1
mydata4.df$ethnic2[mydata4.df$a58=="indian"] <-3
mydata4.df$ethnic2[mydata4.df$a58=="pakistani"] <-6
mydata4.df$ethnic2[mydata4.df$a58=="bangladeshi"] <-5
mydata4.df$ethnic2[mydata4.df$a58=="chin!ese"] <-2
mydata4.df$ethnic2[mydata4.df$a58=="other asian"] <-7
mydata4.df$ethnic2[mydata4.df$a58=="any other"] <-7
mydata4.df$ethnic2[mydata4.df$a58=="mixed ethnic origin"] <-7
mydata4.df$ethnic2[mydata4.df$a58=="ref!used"] <-7
```

In [21]:

```
table(mydata4.df$ethnic2)
```

Out[21]:

```
    1     2     3     4     5     6 
  180    67   369 11962    52   159 
```

Just to check the old variable "ethnic1" and the new variable "ethnic2".

In [22]:

```
mytable <- table (mydata4.df$ethnic1,mydata4.df$ethnic2)
mytable # print table
```

Out[22]:

```
   
        1     2     3     4     5     6
  1     0     0     0 11962     0     0
  2   180     0     0     0     0     0
  3     0     0   369     0     0     0
  4     0     0     0     0     0   159
  5     0     0     0     0    52     0
  6     0    67     0     0     0     0
```

I will now try to re-estimate the model but with the ethnicity variable "ethnic2".

The data have been altered so I re-set the survey design.

In [23]:

In [23]:

```
mydesign3 <- svydesign(id = ~serial,data = mydata4.df, weight = ~weight)
```

In [24]:

```
model4<-svyglm (s15a_c ~ factor(ethnic2) + girls + prof_man + o_non_man + 
                 skilled_man + semi_skilled, design=mydesign3, data = mydata
4.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

In [25]:

```
summary(model4)
```

Out[25]:

```

Call:
svyglm(formula = s15a_c ~ factor(ethnic2) + girls + prof_man + 
    o_non_man + skilled_man + semi_skilled, design = mydesign3, 
    data = mydata4.df, family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.20829    0.19802 -11.152  < 2e-16 ***
factor(ethnic2)2  2.00231    0.37734   5.306 1.14e-07 ***
factor(ethnic2)3  1.06584    0.20829   5.117 3.15e-07 ***
factor(ethnic2)4  0.64314    0.17118   3.757 0.000173 ***
factor(ethnic2)5  0.76616    0.34486   2.222 0.026323 *  
factor(ethnic2)6  0.53136    0.24503   2.169 0.030135 *  
girls             0.40456    0.03926  10.305  < 2e-16 ***
prof_man          2.19209    0.10863  20.179  < 2e-16 ***
o_non_man         1.77251    0.10793  16.423  < 2e-16 ***
skilled_man       0.93217    0.10411   8.954  < 2e-16 ***
semi_skilled      0.57587    0.11264   5.112 3.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000269)

Number of Fisher Scoring iterations: 4
```

Note!

The results now duplicate Table 5 (p.20) Connolley (2006).

I now past the modelling results to the quasi-variance estimation package.

In [26]:

```
model4.qvs <- qvcalc(model4, "factor(ethnic2)")
```

I now get a summary of these results.

This includes the
parameter estimate (i.e. beta) "estimate";
conventional standard error "SE";
quasi-variance based standard error "quasiSE";
quasi-varianve based variance "quasiVar".

In [27]:

```
summary(model4.qvs, digits = 4)
```

```
Model call:  svyglm(formula = s15a_c ~ factor(ethnic2) + girls + prof_man +
o_non_man + skilled_man + semi_skilled, design = mydesign3,      data = myd
ata4.df, family = "binomial")
Factor name:  factor(ethnic2)
      estimate     SE quasiSE  quasiVar
    1   0.0000 0.0000 0.17011 0.0289357
    2   2.0023 0.3773 0.33645 0.1131998
    3   1.0658 0.2083 0.12014 0.0144326
    4   0.6431 0.1712 0.02034 0.0004138
    5   0.7662 0.3449 0.29977 0.0898622
    6   0.5314 0.2450 0.17612 0.0310191
Worst relative errors in SEs of simple contrasts (%):  -0.1 0.1
Worst relative errors over *all* contrasts (%):  -0.4 0.1
```

I now plot the estimates for "ethnicity2" along with quasi-variance based 95% comparison intervals.

In [28]:

```
plot(model4.qvs)
```

The levels for factor(ethnic2)

1 Black; 2 Chinese; 3 Indian; 4 White; 5 Bangladeshi; 6 Pakistani.

Comments

My suspicion that the 'Black' pupils category is a sub-optimal reference category is confirmed.

This is a small category (*n*=180), and there is a large comparison interval around this estimate.

Also, whilst the other five ethnic categories are significantly different to zero (when Black pupils are set as the reference category) the differences between some categories are not significant. For a fuller discussion of using quasi-variance comparison intervals see Gayle and Lambert 2007.

I will now re-estimate the model with 'White' pupils as the referene category.

I will re-organise the ethnic categories as follows

'White'
'Chinese'

'Chinese'

Then the three South Asian categories

'Indian'
'Banglasdeshi'
'Pakistani'

Then finally...

'Black'

(Others are absent from the model)

Re-ordering the ethnicity variable "ethnic2".

Creating a new ethnicity variable "ethnic3".

In [29]:

```
# create the new field,

mydata4.df$ethnic3 <- 7

# everyone is placed into category 7.

# recode the new field with values from the old field.
mydata4.df$ethnic3[mydata4.df$a58=="white"] <-1
mydata4.df$ethnic3[mydata4.df$a58=="carib."] <-6
mydata4.df$ethnic3[mydata4.df$a58=="afro."] <-6
mydata4.df$ethnic3[mydata4.df$a58=="other black"] <-6
mydata4.df$ethnic3[mydata4.df$a58=="indian"] <-3
mydata4.df$ethnic3[mydata4.df$a58=="pakistani"] <-5
mydata4.df$ethnic3[mydata4.df$a58=="bangladeshi"] <-4
mydata4.df$ethnic3[mydata4.df$a58=="chin!ese"] <-2
mydata4.df$ethnic3[mydata4.df$a58=="other asian"] <-7
mydata4.df$ethnic3[mydata4.df$a58=="any other"] <-7
mydata4.df$ethnic3[mydata4.df$a58=="mixed ethnic origin"] <-7
mydata4.df$ethnic3[mydata4.df$a58=="ref!used"] <-7
```

In [30]:

```
table(mydata4.df$ethnic3)
```

Out[30]:

```
    1     2     3     4     5     6 
11962    67   369    52   159   180 
```

Just to check the old variable "ethnic1" and the new variable "ethnic3".

In [31]:

```
mytable <- table (mydata4.df$ethnic1,mydata4.df$ethnic3)
mytable # print table
```

Out[31]:

```
        1     2     3     4     5     6
  1 11962     0     0     0     0     0
  2     0     0     0     0     0   180
  3     0     0   369     0     0     0
  4     0     0     0     0   159     0
  5     0     0     0    52     0     0
  6     0    67     0     0     0     0
```

This looks satisfactory.

I will now try to re-estimate the model but with the ethnicity variable "ethnic2".

The data have been altered so I re-set the survey design.

In [32]:

```
mydesign4 <- svydesign(id = ~serial,data = mydata4.df, weight = ~weight)
```

In [33]:

```
model5<-svyglm (s15a_c ~ factor(ethnic3) + girls + prof_man + o_non_man + 
                skilled_man + semi_skilled, design=mydesign4, data = mydata
4.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

In [34]:

```
summary(model5)
```

Out[34]:

```
Call:
svyglm(formula = s15a_c ~ factor(ethnic3) + girls + prof_man + 
    o_non_man + skilled_man + semi_skilled, design = mydesign4, 
    data = mydata4.df, family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata4.df, weight = ~weight)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.56515    0.10196 -15.351  < 2e-16 ***
factor(ethnic3)2  1.35917    0.33721   4.031 5.59e-05 ***
factor(ethnic3)3  0.42271    0.12187   3.469 0.000525 ***
factor(ethnic3)4  0.12302    0.30052   0.409 0.682278    
factor(ethnic3)5 -0.11177    0.17735  -0.630 0.528544    
factor(ethnic3)6 -0.64314    0.17118  -3.757 0.000173 ***
girls             0.40456    0.03926  10.305  < 2e-16 ***
prof_man          2.19209    0.10863  20.179  < 2e-16 ***
o_non_man         1.77251    0.10793  16.423  < 2e-16 ***
skilled_man       0.93217    0.10411   8.954  < 2e-16 ***
semi_skilled      0.57587    0.11264   5.112 3.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1.000269)

Number of Fisher Scoring iterations: 4
```

Note!

The results no longer **duplicate Table 5 (p.20) Connolley (2006)** but the results build upon, and extend, the work so they are a **replication**.

I now pass the results to the quasi-variance procedure.

In [35]:

```
model5.qvs <- qvcalc(model5, "factor(ethnic3)")
```

In [36]:

```
summary(model5.qvs, digits = 4)
```

```
Model call:  svyglm(formula = s15a_c ~ factor(ethnic3) + girls + prof_man +
o_non_man + skilled_man + semi_skilled, design = mydesign4,      data = myd
ata4.df, family = "binomial")
Factor name:  factor(ethnic3)
      estimate     SE quasiSE  quasiVar
    1   0.0000 0.0000 0.02034 0.0004138
    2   1.3592 0.3372 0.33645 0.1131998
    3   0.4227 0.1219 0.12014 0.0144326
    4   0.1230 0.3005 0.29977 0.0898622
    5  -0.1118 0.1773 0.17612 0.0310191
    6  -0.6431 0.1712 0.17011 0.0289357
Worst relative errors in SEs of simple contrasts (%):  -0.1 0.1
Worst relative errors over *all* contrasts (%):  -0.4 0.1
```

I now plot the results for "ethnicity3" along with quasi-variance based 95% comparison intervals.

In [37]:

```
plot(model5.qvs)
```

The levels for factor(ethnic3)

1. White
2. Chinese
3. Indian
4. Banglasdeshi
5. Pakistani
6. Black

(Others are absent from the model)

Comments

The results have been **duplicated** and then built upon. These results are a **replication**.

The model is improved by using 'White' pupils as the reference category. This is a large category and there is a small comparison interval around the estimate.

The model also tells a more theoretically useful substantive story. The use of quasi-variance based comparison intervals allows a more subtle investigation of the differences between ethnic groups.

There are some ethnic differences in school GCSE outcomes (5+ GCSEs at grades A - C).

Compared with the majority of pupils who are white those who are Chinese have better outcomes.

Indian pupils perform better than white pupils.

Bangladeshi and Pakistani pupils do have significantly different outcomes to white pupils.

Black pupils have significantly poorer outcomes than their white counterparts.

It is notable that the three south Asian ethnic groups are not significantly different to each other.

This hopefully illustrates that this model is better parameterised that the original model presented in Table 5 (p.20) Connolly (2006).

The analyses above have required some more data wrangling. Therefore it is prudent to save a new copy of the data.

I will take a look at the objects that are knocking around.

In [39]:

```
ls()
```

Out[39]:

```
"model2"  "model3"  "model4"  "model4.qvs"  "model5"  "model5.qvs"  "mydata4.df"
"mydesign2"  "mydesign3"  "mydesign4"  "mytable"
```

To avoid confusion later and to help to keep the workflow clear I will create a new data frame "mydata5.df".

In [40]:

```
mydata5.df<-mydata4.df
str(mydata5.df)
```

```
'data.frame':	12789 obs. of  27 variables:
 $ serial       : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight       : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex          : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c        : int  9 9 9 9 2 2 7 3 10 1
 $ s1a_c        : int  5 5 5 5 2 2 7 5 10 1 ...
 $ a58          : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth        : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe       : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man     : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man    : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man  : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c       : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls        : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1      : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white        : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian       : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ethnic2      : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3      : num  1 1 3 1 1 1 1 1 1 1 ...
```

In [41]:

```
save(mydata5.df,file="C:/Users/Vernon/OneDrive - University of Edinburgh/Do
cuments/ycs_9_2017/ycs9sw1_v3.rda")
```

Here I make a Stata copy of the file just in case I required it for *swivel chair* activities later in the workflow.

In [ ]:

```
write.dta(mydata5.df, "C:/Users/Vernon/OneDrive - University of Edinburgh/D
ocuments/ycs_9_2017/ycs9sw1_v3.dta")
```

Replicating the Connolly (2006) Model Results Adding an Improved Social Class Measure (UK National Socio-economic Classification - NS-SEC)

In this next stage of the analysis I will explore importing an additions social class measure.

The measure of social class that is employed in Table 5 (p.20) Connelly (2006) is unconventional in social stratification research.

The National Socio-economic Classification (NS-SEC) is a commonly used measure in stratification research and is the measure used in official statistics and government research in the United Kingdom.

In the next stage of the analysis I replicate the analysis of school GCSE attainment using YCS Cohort 9 through the incorporation of a parental NS-SEC measure that was derrived by Croxford et al (2007).

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

Warning.

Within this Jupyter notebook there has been a lot of non-routine work. For example I have 'swivel-chaired' between data analytical software packages and changed kernels.

It may from time to time be necessary to re-start the notebook depending on how stable your computing environment is.

In this section I re-start a *R* session.

In [11]:

```
library(foreign)
library(survey)
library(car)
library(dplyr)
library(weights)
library(dummies)

library(MASS)
library(qvcalc)
```

```
Warning message:
```

```
: package 'weights' was built under R version 3.2.5Loading required package
: Hmisc
Warning message:
: package 'Hmisc' was built under R version 3.2.5Loading required package:
ggplot2
Warning message:
: package 'ggplot2' was built under R version 3.2.5
```

```
Error: package 'ggplot2' could not be loaded
```

Various WARNINGS will appear. Don't panic.

From the Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 UK Data Archive Study 5765, I will import a file called "ew_core". This is the core file containing pupils in England and Wales.

In [2]:

```
# This file is located on (my) OneDrive.

mydataew.df <- read.dta("C:/Users/Vernon/OneDrive - University of Edinburgh
/Documents/ycs_9_2017/ew_core.dta")
```

In [3]:

```
summary(mydataew.df)
```

Out[3]:

```
    t0cohort        t0nation          t0caseid           t0source        
 Min.   :1984   england :107922   Min.   :   100001   Length:115179     
 1st Qu.:1988   wales   :  7257   1st Qu.:   131432   Class :character  
 Median :1993   scotland:     0   Median :228404103   Mode  :character  
 Mean   :1992                     Mean   :339926553                     
 3rd Qu.:1995                     3rd Qu.:680400520                     
 Max.   :1999                     Max.   :996602914                     

    t1weight         t2weight        t3weight                  t1resp      
 Min.   :0.1011   Min.   :-1.00   Min.   :-1.00   did not respond:     0  
 1st Qu.:0.7269   1st Qu.: 0.63   1st Qu.: 0.53   respondent     :115179  
 Median :0.9122   Median : 0.82   Median : 0.72                           
 Mean   :1.0000   Mean   : 0.92   Mean   : 0.83                           
 3rd Qu.:1.1777   3rd Qu.: 1.15   3rd Qu.: 1.07                           
 Max.   :3.8550   Max.   : 6.09   Max.   : 8.29                           
                  NA's   :63671   NA's   :42871                           
             t2resp                   t3resp         t0schtyp      
 no survey at t2:47618   did not respond:52334   Min.   :  1.000  
 did not respond:18042   respondent     :62845   1st Qu.:  2.000  
 respondent     :49519                           Median :  3.000  
                                                 Mean   :  3.195  
                                                 3rd Qu.:  3.000  
                                                 Max.   :999.000  

                t0sex                         t0stay             t0sibs      
 not answered (9)   :    0   not answered     : 1481   Min.   :-9.000  
 item not applicable:    0   none             :    0   1st Qu.: 1.000  
 male               :54396   father and mother:93404   Median : 1.000  
 female             :60783   mother only      :14220   Mean   : 1.364  
                             father only      : 3022   3rd Qu.: 2.000  
                             other response   : 3052   Max.   :23.000  
```

```
                            other response   : 5052   Max.   :25.000

          t0ethnic              t0house                t0dadpce
white         :101695   not answered: 2137   not answered  :10847
indian        :  2612   owned       :78293   yes           :16920
not answered  :  2046   rented      :18826   no            :34913
survey problem:  1989   other       : 1807   other response:14111
pakistani     :  1738   NA's        :14116   NA's          :38388
other response:  1715
(Other)       :  3384
          t0mumpce                t0dadalv                t0mumalv
not answered  : 8624   not answered  :10847   not answered  : 8624
yes           :15815   yes           :16920   yes           :15815
no            :38718   no            :34913   no            :38718
other response:13634   other response:14111   other response:13634
NA's          :38388   NA's          :38388   NA's          :38388


          t0daddeg                t0mumdeg                     t0dadjob
not answered  :13278   not answered  :11654   not answered (9)   : 9867
yes           :12862   yes           : 8826   item not applicable:    0
no            :39142   no            :44957   yes                :89865
other response:11509   other response:11354   no                 :15447
NA's          :38388   NA's          :38388


               t0mumjob                           t0truant
not answered (9)   : 6003   not answered              : 1539
item not applicable:    0   weeks at a time           : 2025
yes                :55195   days at a time            : 2473
no                 :53981   occasional days or lessons:43961
                            never                     :65181


                t1att1                      t1att2
not answered (9)   : 3241   not answered (9)   : 3588
item not applicable:    0   item not applicable:    0
agree              :62450   agree              :40737
disagree           :33589   disagree           :54955
NA's               :15899   NA's               :15899


                t1att3                  t0region        t0dadsoc
not answered (9)   : 3658   other south east:24608   Min.   : -9.0
item not applicable:    0   north west      :14674   1st Qu.:126.0
agree              :61018   west midlands   :12439   Median :331.0
disagree           :34604   yorks & humber  :11942   Mean   :375.5
NA's               :15899   greater london  :11528   3rd Qu.:570.0
                            (Other)         :39986   Max.   :999.0
                            NA's            :    2   NA's   :30324
   t0mumsoc        t0examst          t0examac          t0examaf
Min.   : -9.0   Min.   :-9.000   Min.   :-9.000   Min.   :-9.000
1st Qu.: -9.0   1st Qu.: 7.000   1st Qu.: 1.000   1st Qu.: 6.000
Median :390.0   Median : 8.000   Median : 4.000   Median : 8.000
Mean   :374.3   Mean   : 7.906   Mean   : 4.256   Mean   : 6.918
3rd Qu.:644.0   3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.: 9.000
Max.   :999.0   Max.   :18.000   Max.   :16.000   Max.   :16.000
NA's   :30324
   t0score          t0vocsbj       t0vocpas        t0othsbj
Min.   : -9.00   Min.   :0.00   Min.   :0.00   Min.   :0.0000
1st Qu.: 21.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.0000
```

```
 Median : 36.00   Median :0.00   Median :0.00   Median :0.0000  
 Mean   : 34.77   Mean   :0.13   Mean   :0.12   Mean   :0.1138  
 3rd Qu.: 50.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.0000  
 Max.   :112.00   Max.   :8.00   Max.   :8.00   Max.   :8.0000  
                  NA's   :8064   NA's   :8064                   
    t0othpas           t1dooct           t1donow            t0age      
 Min.   :0.00000   Min.   :-9.000   Min.   :-9.000   Min.   :-9.00  
 1st Qu.:0.00000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:16.00  
 Median :0.00000   Median : 3.000   Median : 3.000   Median :16.25  
 Mean   :0.09774   Mean   : 3.049   Mean   : 3.176   Mean   :15.85  
 3rd Qu.:0.00000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.:16.50  
 Max.   :8.00000   Max.   :10.000   Max.   :10.000   Max.   :16.75  
                                                     NA's   :3      
                t0dadse                       t0mumse     
 not answered (9)   :14725   not answered (9)   :21914  
 item not applicable: 3889   item not applicable: 3889  
 yes                :23196   yes                : 8972  
 no                 :73369   no                 :80404  




                     t0gor                              t0urban     
 south east             :24602   not urban (lt 90%)         :57844  
 north west             :16553   urban area not in top 10:26023  
 west midlands          :12439   greater london             :11528  
 yorkshire & humberside:11648   west midlands ua           : 5417  
 london                 :11528   greater manchester         : 5051  
 (Other)                :38407   (Other)                    : 9314  
 NA's                   :    2   NA's                       :    2  
    t0mumsec         t0dadsec         t0parsec         t0dadsc4    
 Min.   : 1.10   Min.   : 1.10   Min.   : 1.10   Min.   :  1.0  
 1st Qu.: 3.00   1st Qu.: 2.00   1st Qu.: 2.00   1st Qu.:  1.0  
 Median : 6.00   Median : 4.00   Median : 3.00   Median :  2.0  
 Mean   :30.04   Mean   :22.26   Mean   :14.03   Mean   : 54.1  
 3rd Qu.:99.00   3rd Qu.: 7.00   3rd Qu.: 6.00   3rd Qu.:  3.0  
 Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :999.0  
 NA's   :38388   NA's   :38388   NA's   :38388                   
    t0mumsc4          t0parsc4          t0monthb         t3alev     
 Min.   :  1.00   Min.   :  1.00   Min.   : 0.000   no  :84329  
 1st Qu.:  2.00   1st Qu.:  1.00   1st Qu.: 3.000   yes :22786  
 Median :  3.00   Median :  2.00   Median : 6.000   NA's: 8064  
 Mean   : 62.62   Mean   : 45.91   Mean   : 6.332               
 3rd Qu.: 99.00   3rd Qu.:  3.00   3rd Qu.: 9.000               
 Max.   :999.00   Max.   :999.00   Max.   :99.000               
                                   NA's   :2                    
    t3nqf_a          t3_ucas          t3uscore          t3lev3       t3twoa     
 Min.   :0.000   Min.   :  0.00   Min.   :   0.00   no  :83834   no  :87006  

 1st Qu.:0.000   1st Qu.:  0.00   1st Qu.:   0.00   yes :23281   yes :20109  

 Median :0.000   Median :  0.00   Median :   0.00   NA's: 8064   NA's: 8064  
 Mean   :0.765   Mean   : 53.03   Mean   :  57.91                             
 3rd Qu.:2.000   3rd Qu.:  0.00   3rd Qu.:  38.00                             

 Max.   :3.000   Max.   :990.00   Max.   :1008.00                             
 NA's   :8064    NA's   :8064     NA's   :8064                                
    t3nowed                                                 t3nowhe     
 Min.   :-9.0    missing information                           : 1539  
 1st Qu.: 0.0    no, in he but in other non-advanced cources:44591  
 Median : 0.0    yes, in he                                    :16715  
 Mean   : 0.4    NA's                                        :52334  
 3rd Qu.: 1.0                                                       
 Max.   : 1.0                                                       
 NA's   :52334                                                      
                                                  t3degree     
 missing information                                  :    0  
 no, studying for a non-advanced qualification        :46080  
 yes, studying for a degree                           :15226  
 no, studying for another advanced non-university:    0  
 NA's                                                 :53873  
                                                                    
                                                                    
                         t3dooct                                   t3donow     
 full time education           :24817   full time education           :27674  
 government supported training: 2385   full time job                 :23202  
 full time job                 :22145    unemployed                   : 4244  
  unemployed                   : 3778   government supported training: 2229  
 something else                : 7070   part-time job                 : 2206  
 NA's                          :54984   (Other)                       : 2956  
                                        NA's                          :52668  
                         t2dooct                                   t2donow     
 full time education           :43886   full time education           :42880  
 government supported training: 6084   full time job                 :12702  
 full time job                 :11721   government supported training: 5549  
  unemployed                   : 1907    unemployed                   : 2354  
 something else                : 1481   part-time job                 : 1097  
 NA's                          :50100   (Other)                       :  888  
                                        NA's                          :49709  
```

To see the summary of the data use the scroll bar to **scroll down**.

In [4]:

```
str(mydataew.df)
```

```
'data.frame':	115179 obs. of  66 variables:
 $ t0cohort: int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
 $ t0nation: Factor w/ 3 levels "england","wales",..: 1 1 1 1 1 1 1 1 1 1
...
 $ t0caseid: int  301402257 301402259 301402260 301402652 301402654 3014026
56 301402660 301402661 301402666 301402668 ...
 $ t0source: chr  "ycs1" "ycs1" "ycs1" "ycs1" ...
 $ t1weight: num  1.25 2.1 1.05 1.61 2.1 ...
 $ t2weight: num  -1 3.732 0.866 -1 -1 ...
 $ t3weight: num  -1 3.254 0.802 -1 -1 ...
 $ t1resp  : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2 2
...
 $ t2resp  : Factor w/ 3 levels "no survey at t2",..: 2 3 3 2 2 2 3 3 2 2
...
 $ t3resp  : Factor w/ 2 levels "did not respond",..: 1 2 2 1 1 1 1 1 1 1
...
 $ t0schtyp: int  3 3 3 3 3 3 3 3 3 3 ...
 $ t0sex   : Factor w/ 4 levels "not answered (9)",..: 4 3 4 3 3 4 4 4 3 4
...
 $ t0stay  : Factor w/ 6 levels "not answered",..: 6 3 3 6 4 3 3 6 3 6 ...
 $ t0sibs  : int  2 2 1 1 0 1 3 2 2 7 ...
 $ t0ethnic: Factor w/ 9 levels "not answered",..: 2 3 3 2 2 2 9 3 2 2 ...
 $ t0house : Factor w/ 4 levels "not answered",..: 1 2 2 1 3 3 2 2 3 3 ...
 $ t0dadpce: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
 $ t0dadpce: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0mumpce: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0dadalv: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0mumalv: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0daddeg: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0mumdeg: Factor w/ 4 levels "not answered",..: NA NA NA NA NA NA NA NA
NA NA ...
 $ t0dadjob: Factor w/ 4 levels "not answered (9)",..: 3 3 3 1 1 3 3 3 3 3
...
 $ t0mumjob: Factor w/ 4 levels "not answered (9)",..: 4 4 3 1 3 4 4 4 4 4
...
 $ t0truant: Factor w/ 5 levels "not answered",..: 4 4 5 5 4 4 3 5 5 5 ...
 $ t1att1  : Factor w/ 4 levels "not answered (9)",..: 4 3 3 4 4 4 3 3 3 3
...
 $ t1att2  : Factor w/ 4 levels "not answered (9)",..: 4 4 3 3 3 3 4 4 4 4
...
 $ t1att3  : Factor w/ 4 levels "not answered (9)",..: 3 3 3 4 4 4 3 4 4 3
...
 $ t0region: Factor w/ 14 levels "not answered (99)",..: 11 11 11 11 11 11
11 11 11 11 ...
 $ t0dadsoc: int  -9 -9 615 -9 -9 620 535 872 532 872 ...
 $ t0mumsoc: int  -9 -9 460 -9 941 722 553 -9 620 958 ...
 $ t0examst: int  6 6 8 0 7 7 6 7 8 7 ...
 $ t0examac: int  0 0 6 0 0 0 0 1 0 1 ...
 $ t0examaf: int  3 0 7 0 2 3 1 5 3 5 ...
 $ t0score : int  10 0 34 0 7 9 4 17 10 19 ...
 $ t0vocsbj: int  NA NA NA NA NA NA NA NA NA NA ...
 $ t0vocpas: int  NA NA NA NA NA NA NA NA NA NA ...
 $ t0othsbj: int  0 0 2 0 0 0 0 1 0 2 ...
 $ t0othpas: int  0 0 1 0 0 0 0 0 0 2 ...
 $ t1dooct : int  5 6 6 -9 6 6 6 5 5 5 ...
 $ t1donow : int  5 6 6 -9 6 6 6 5 5 6 ...
 $ t0age   : num  -9 15.8 16.5 -9 -9 ...
 $ t0dadse : Factor w/ 4 levels "not answered (9)",..: 1 1 4 1 1 4 4 4 4 4
...
 $ t0mumse : Factor w/ 4 levels "not answered (9)",..: 1 1 4 1 4 4 4 1 4 4
...
 $ t0gor   : Factor w/ 11 levels "north east","north west",..: 7 7 7 7 7 7
7 7 7 7 ...
 $ t0urban : Factor w/ 12 levels "not urban (lt 90%)",..: 2 2 2 2 2 2 2 2 2
2 ...
 $ t0mumsec: num  NA NA NA NA NA NA NA NA NA NA ...
 $ t0dadsec: num  NA NA NA NA NA NA NA NA NA NA ...
 $ t0parsec: num  NA NA NA NA NA NA NA NA NA NA ...
 $ t0dadsc4: int  99 99 3 99 99 3 3 3 3 3 ...
 $ t0mumsc4: int  99 99 2 99 3 2 3 99 3 3 ...
 $ t0parsc4: int  99 99 2 99 3 2 3 3 3 3 ...
 $ t0monthb: int  0 8 12 0 0 0 2 4 0 0 ...
 $ t3alev  : Factor w/ 2 levels "no","yes": NA NA NA NA NA NA NA NA NA NA .
..
 $ t3nqf_a : num  NA NA NA NA NA NA NA NA NA NA ...
 $ t3_ucas : int  NA NA NA NA NA NA NA NA NA NA ...
 $ t3uscore: int  NA NA NA NA NA NA NA NA NA NA ...
 $ t3lev3  : Factor w/ 2 levels "no","yes": NA NA NA NA NA NA NA NA NA NA .
..
 $ t3twoa  : Factor w/ 2 levels "no","yes": NA NA NA NA NA NA NA NA NA NA .
```

```
..
 $ t3nowed : int  NA 0 0 NA NA NA NA NA NA NA ...
 $ t3nowhe : Factor w/ 3 levels "missing information",..: NA 2 2 NA NA NA N
A NA NA NA ...
 $ t3degree: Factor w/ 4 levels "missing information",..: NA 2 2 NA NA NA N
A NA NA NA ...
 $ t3dooct : Factor w/ 5 levels "full time education",..: NA 3 3 NA NA NA N
A NA NA NA ...
 $ t3donow : Factor w/ 7 levels "full time education",..: NA 3 3 NA NA NA N
A NA NA NA ...
 $ t2dooct : Factor w/ 5 levels "full time education",..: NA 3 3 NA NA NA N
A NA NA NA ...
 $ t2donow : Factor w/ 7 levels "full time education",..: NA 3 3 NA NA NA N
A NA NA NA ...
 - attr(*, "datalabel")= chr ""
 - attr(*, "time.stamp")= chr ""
 - attr(*, "formats")= chr  "%8.0g" "%8.0g" "%12.0g" "%5s" ...
 - attr(*, "types")= int  252 251 253 5 255 255 255 251 251 251 ...
 - attr(*, "val.labels")= chr  "" "t0nation" "" "" ...
 - attr(*, "var.labels")= chr  "year completed compulsory schooling" "natio
nal system" "id for time series" "source of data" ...
 - attr(*, "version")= int 8
 - attr(*, "label.table")=List of 47
  ..$ t0nation: Named int  1 2 3
  .. ..- attr(*, "names")= chr  "england" "wales" "scotland"
  ..$ t1resp  : Named int  0 1
  .. ..- attr(*, "names")= chr  "did not respond" "respondent"
  ..$ t2resp  : Named int  -1 0 1
  .. ..- attr(*, "names")= chr  "no survey at t2" "did not respond" "respon
dent"
  ..$ t3resp  : Named int  0 1
  .. ..- attr(*, "names")= chr  "did not respond" "respondent"
  ..$ t0schtyp: Named int  1 2 3 4 5 6 7 8
  .. ..- attr(*, "names")= chr  "6th form college" "comp to 16" "comp to 18
" "grammar" ...
  ..$ t0sex   : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "m
ale" "female"
  ..$ t0stay  : Named int  -9 0 1 2 3 4
  .. ..- attr(*, "names")= chr  "not answered" "none" "father and mother" "
mother only" ...
  ..$ t0ethnic: Named int  -9 -1 1 4 5 6 7 9 10
  .. ..- attr(*, "names")= chr  "not answered" "survey problem" "white" "bl
ack" ...
  ..$ t0house : Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "owned" "rented" "other"
  ..$ t0dadpce: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0mumpce: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0dadalv: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0mumalv: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0daddeg: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0mumdeg: Named int  -9 1 2 3
  .. ..- attr(*, "names")= chr  "not answered" "yes" "no" "other response"
  ..$ t0dadjob: Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "y
```

```
es" "no"
  ..$ t0mumjob: Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "y
es" "no"
  ..$ t0truant: Named int  -9 1 2 3 4
  .. ..- attr(*, "names")= chr  "not answered" "weeks at a time" "days at a
time" "occasional days or lessons" ...
  ..$ t1att1  : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "a
gree" "disagree"
  ..$ t1att2  : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "a
gree" "disagree"
  ..$ t1att3  : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "a
gree" "disagree"
  ..$ t0region: Named int  -9 -6 -2 -1 1 2 3 4 5 6 ...
  .. ..- attr(*, "names")= chr  "not answered (99)" "schedule not obtained"
"schedule not applicable" "item not applicable" ...
  ..$ t0dadsoc: Named int  100 101 102 103 110 111 112 113 120 121 ...
  .. ..- attr(*, "names")= chr  "asst secty nat govt     " "company gen mana
ger    " "local govt officers    " "heo\\sen prl nat gov    " ...
  ..$ t0mumsoc: Named int  100 101 102 103 110 111 112 113 120 121 ...
  .. ..- attr(*, "names")= chr  "asst secty nat govt     " "company gen mana
ger    " "local govt officers    " "heo\\sen prl nat gov    " ...
  ..$ t1dooct : Named int  0 1 2 3 4 5 6 7 8 9 ...
  .. ..- attr(*, "names")= chr  "nk" "school" "6th form college" "fe colleg
e" ...
  ..$ t1donow : Named int  0 1 2 3 4 5 6 7 8 9 ...
  .. ..- attr(*, "names")= chr  "nk" "school" "6th form college" "fe colleg
e" ...
  ..$ t0dadse : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "y
es" "no"
  ..$ t0mumse : Named int  -9 -1 1 2
  .. ..- attr(*, "names")= chr  "not answered (9)" "item not applicable" "y
es" "no"
  ..$ t0gor   : Named int  1 2 3 4 5 6 7 8 9 10 ...
  .. ..- attr(*, "names")= chr  "north east" "north west" "yorkshire & humb
erside" "east midlands" ...
  ..$ t0urban : Named int  0 1 2 3 4 5 6 7 8 9 ...
  .. ..- attr(*, "names")= chr  "not urban (lt 90%)" "greater london" "west
midlands ua" "greater manchester" ...
  ..$ t0mumsec: Named int  1 1 2 3 4 5 6 7 99
  .. ..- attr(*, "names")= chr  "higher managerial" "professional" "lower m
anagerial and professional" "intermediate" ...
  ..$ t0dadsec: Named int  1 1 2 3 4 5 6 7 99
  .. ..- attr(*, "names")= chr  "higher managerial" "professional" "lower m
anagerial and professional" "intermediate" ...
  ..$ t0parsec: Named int  1 1 2 3 4 5 6 7 99
  .. ..- attr(*, "names")= chr  "higher managerial" "professional" "lower m
anagerial and professional" "intermediate" ...
  ..$ t0dadsc4: Named int  1 2 3 99
  .. ..- attr(*, "names")= chr  "managerial & professional" "intermediate"
"working" "unclassified"
  ..$ t0mumsc4: Named int  1 2 3 99
  .. ..- attr(*, "names")= chr  "managerial & professional" "intermediate"
"working" "unclassified"
  ..$ t0parsc4: Named int  1 2 3 99
  .. ..- attr(*, "names")= chr  "managerial & professional" "intermediate"
```

```
"working" "unclassified"
  ..$ t0monthb: Named int  1 2 3 4 5 6 7 8 9 10 ...
  .. ..- attr(*, "names")= chr  "january" "february" "march" "april" ...
  ..$ t3alev  : Named int  0 1
  .. ..- attr(*, "names")= chr  "no" "yes"
  ..$ t3lev3  : Named int  0 1
  .. ..- attr(*, "names")= chr  "no" "yes"
  ..$ t3twoa  : Named int  0 1
  .. ..- attr(*, "names")= chr  "no" "yes"
  ..$ t3nowed : Named int  0 1
  .. ..- attr(*, "names")= chr  "in other status" "in full-time education"
  ..$ t3nowhe : Named int  -9 0 1
  .. ..- attr(*, "names")= chr  "missing information" "no, in he but in oth
er non-advanced cources" "yes, in he"
  ..$ t3degree: Named int  -9 0 1 2
  .. ..- attr(*, "names")= chr  "missing information" "no, studying for a n
on-advanced qualification" "yes, studying for a degree" "no, studying for a
nother advanced non-university"
  ..$ t3dooct : Named int  1 5 6 8 10
  .. ..- attr(*, "names")= chr  "full time education" "government supported
training" "full time job" " unemployed" ...
  ..$ t3donow : Named int  1 5 6 7 8 9 10
  .. ..- attr(*, "names")= chr  "full time education" "government supported
training" "full time job" "part-time job" ...
  ..$ t2dooct : Named int  1 5 6 8 10
  .. ..- attr(*, "names")= chr  "full time education" "government supported
training" "full time job" " unemployed" ...
  ..$ t2donow : Named int  1 5 6 7 8 9 10
  .. ..- attr(*, "names")= chr  "full time education" "government supported
training" "full time job" "part-time job" ...
```

The variables that I require are

t0cohort - the YCS cohort (i.e. year).

t0nation - identifies if the pupil is from the England and Wales data (this is just a check the dataset should be England and Wales on hence "ew" in "ew_core.dta" .

t0caseid - this is an *id* variable. However, it is not unqiue across YCS cohorts so **must** be used in conjuction with a **cohort** identifier.

t0source - identifies the YCS cohort (e.g. YCS 9).

t1weight - this is the sweep 1 survey weight.

t1resp - identifies if the pupil responded in sweep 1 of the survey.

t0parsec - this is the parental NS-SEC measure (8 category) that is derrived by Croxford et al. (2007). This is the measure that I require for the current replication exercise.

In [5]:

```
table(mydataew.df$t0cohort)
table(mydataew.df$t0source)
table(mydataew.df$t0parsec)
```

Out[5]:

```
 1984  1986  1988  1990  1993  1995  1997  1999
 8064 16208 14116 14511 18021 15899 14662 13698
```

Out[5]:

```
 ycs1 ycs10  ycs3  ycs4  ycs5  ycs7  ycs8  ycs9 
 8064 13698 16208 14116 14511 18021 15899 14662 
```

Out[5]:

```
  1.1   1.2     2     3     4     5     6     7    99 
 4533  7807 17171 11518 11349  4055  7335  4398  8625 
```

Get a subset of the data with only the variables needed.

In [6]:

```
myvarsew <- c("t0cohort", "t0nation", "t0caseid", "t0source", "t1weight", "
t1resp", "t0parsec")
mydataew.df <- mydataew.df[myvarsew]
```

In [7]:

```
summary(mydataew.df)
```

Out[7]:

```
    t0cohort         t0nation          t0caseid            t0source        
 Min.   :1984   england :107922   Min.   :   100001   Length:115179     
 1st Qu.:1988   wales   :  7257   1st Qu.:   131432   Class :character  
 Median :1993   scotland:     0   Median :228404103   Mode  :character  
 Mean   :1992                     Mean   :339926553                     
 3rd Qu.:1995                     3rd Qu.:680400520                     
 Max.   :1999                     Max.   :996602914                     
                                                                        
    t1weight                   t1resp           t0parsec    
 Min.   :0.1011   did not respond:     0   Min.   : 1.10  
 1st Qu.:0.7269   respondent     :115179   1st Qu.: 2.00  
 Median :0.9122                            Median : 3.00  
 Mean   :1.0000                            Mean   :14.03  
 3rd Qu.:1.1777                            3rd Qu.: 6.00  
 Max.   :3.8550                            Max.   :99.00  
                                           NA's   :38388  
```

Now I get a subset of the cases (i.e. pupils) that are in the YCS cohort 9.

In [8]:

```
mydataew2.df <- mydataew.df[ which(mydataew.df$t0source=="ycs9"),]
```

In [52]:

```
summary(mydataew2.df)
table(mydataew2.df$t0source)
```

Out[52]:

```
    t0cohort         t0nation         t0caseid          t0source        
 Min.   :1997   england :13762   Min.   :200001   Length:14662      
 1st Qu.:1997   wales   :  900   1st Qu.:206123   Class :character  
 Median :1997   scotland:    0   Median :211589   Mode  :character
 Mean   :1997                    Mean   :212056
 3rd Qu.:1997                    3rd Qu.:217027
 Max.   :1997                    Max.   :231392
    t1weight                  t1resp          t0parsec
 Min.   :0.6025   did not respond:    0   Min.   : 1.10
 1st Qu.:0.7661   respondent     :14662   1st Qu.: 2.00
 Median :0.8779                           Median : 3.00
 Mean   :1.0000                           Mean   :12.99
 3rd Qu.:1.0576                           3rd Qu.: 6.00
 Max.   :2.5176                           Max.   :99.00
```

```
Out[52]:
```

```
 ycs9
14662
```

I will now check which objects are knocking around.

```
In [9]:
```

```
ls()
```

```
Out[9]:
```

"mydataew.df"   "mydataew2.df"   "myvarsew"

I will now (re-)load the last version of my YCS cohort 9 dataset "ycs9sw1_v3.rda".

```
In [10]:
```

```
load("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v3.rda")
ls()
```

```
Out[10]:
```

"mydata5.df"   "mydataew.df"   "mydataew2.df"   "myvarsew"

```
In [55]:
```

```
str(mydata5.df)
```

```
'data.frame':	12789 obs. of  27 variables:
 $ serial       : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight       : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex          : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c        : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58          : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth        : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe       : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man     : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man    : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man  : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c       : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls        : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1      : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white        : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian       : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ethnic2      : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3      : num  1 1 3 1 1 1 1 1 1 1 ...
```

This data frame should have "ethnic2" and "ethnic3" in it.
If they are absent then a older file has been used.

I am now going to wrangle the data a little.

The reshape library is required if it is not already loaded.

In [56]:

```
library(reshape)
```

```
Warning message:
: package 'reshape' was built under R version 3.2.5
Attaching package: 'reshape'

The following object is masked from 'package:dplyr':

    rename

The following object is masked from 'package:Matrix':

    expand
```

The 'id" variables in file "ew_core" (i.e. Croxford's time series files) is not the same as in YCS cohort 9 file "ycs9sw1".
Therefore I am going to change the variable "t0caseid" which is in "ew_core" to "serial" which is the name of the "id" variable in "ycs9sw1".

In [57]:

```
mydataew2.df <- rename(mydataew2.df, c(t0caseid="serial"))
str(mydataew2.df)
```

```
'data.frame': 14662 obs. of  7 variables:
 $ t0cohort: int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 ...
 $ t0nation: Factor w/ 3 levels "england","wales",..: 1 1 1 1 1 1 1 1 1 1
```

```
...
 $ serial  : int  200001 200004 200005 200006 200008 200012 200013 200014 2
00019 200022 ...
 $ t0source: chr  "ycs9" "ycs9" "ycs9" "ycs9" ...
 $ t1weight: num  0.875 0.976 0.976 0.976 1.841 ...
 $ t1resp  : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2 2
...
 $ t0parsec: num  1.1 4 4 2 99 99 99 4 99 99 ...
```

Now I combine the file "ycs9sw1_v3" which is in data frame mydata5.df **with** "ew_core" which is in data frame mydataew2.df.

In [59]:

```
mydata6.df <- merge(mydata5.df, mydataew2.df,by="serial")
```

In [60]:

```
str(mydata6.df)
```

```
'data.frame':	12789 obs. of  33 variables:
 $ serial      : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight      : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex         : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c       : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58         : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth       : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe      : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls       : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1     : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white       : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ethnic2     : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3     : num  1 1 3 1 1 1 1 1 1 1 ...
 $ t0cohort    : int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 ...
 $ t0nation    : Factor w/ 3 levels "england","wales",..: 1 1 1 1 1 1 1 1 1
1 ...
 $ t0source    : chr  "ycs9" "ycs9" "ycs9" "ycs9" ...
 $ t1weight    : num  0.875 0.976 0.976 0.976 0.959 ...
 $ t1resp      : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2
 $ ethresp      : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2
2 ...
 $ t0parsec     : num  1.1 4 4 2 4 1.2 5 1.1 3 4 ...
```

If this has worked then mydata6.df should contain "ethnic2", "ethnic3" and "t0parsec".

Here is the first glimpse at the parental NS-SEC variable "t0parsec".

In [61]:

```
mytablenssec <- table(mydata6.df$t0parsec, mydata6.df$s15a_c)
mytablenssec # print table
```

Out[61]:

```
         0    1
  1.1  163  620
  1.2  215 1151
  2    936 2373
  3    831 1408
  4   1073 1072
  5    433  311
  6    879  538
  7    554  231
  99     1    0
```

In [62]:

```
prop.table (mytablenssec, 1)
```

Out[62]:

```
               0         1
  1.1 0.2081737 0.7918263
  1.2 0.1573939 0.8426061
  2   0.2828649 0.7171351
  3   0.3711478 0.6288522
  4   0.5002331 0.4997669
  5   0.5819892 0.4180108
  6   0.6203246 0.3796754
  7   0.7057325 0.2942675
  99  1.0000000 0.0000000
```

In [63]:

```
save(mydata6.df,file="C:/Users/Vernon/OneDrive - University of Edinburgh/Do
cuments/ycs_9_2017/ycs9sw1_v4.rda")
```

In [28]:

```
load ("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v4.rda")
ls()
```

Out[28]:

"model6" "mydata5.df" "mydata6.df" "mydataew.df" "mydataew2.df" "mydesign5" "mytablenssec" "myvarsew"

Takin a second look at NS-SEC the social class variable.

In [30]:

```
mytablenssec <- table(mydata6.df$t0parsec, mydata6.df$s15a_c)
mytablenssec # print table
```

Out[30]:

```
         0    1
  1.1   163  620
  1.2   215 1151
  2     936 2373
  3     831 1408
  4    1073 1072
  5     433  311
  6     879  538
  7     554  231
  99      1    0
```

There is a missing value coded as "99".

In [31]:

```
mydata6.df$t0parsec[mydata6.df$t0parsec=="99"] <-NA
```

In [32]:

```
mytablenssec <- table(mydata6.df$t0parsec, mydata6.df$s15a_c)
mytablenssec # print table
```

Out[32]:

```
         0    1
  1.1   163  620
  1.2   215 1151
  2     936 2373
  3     831 1408
  4    1073 1072
  5     433  311
  6     879  538
  7     554  231
```

I now check that "t0parsec1" is a factor.

In [33]:

```
levels(mydata6.df$t0parsec )
```

Out[33]:

```
NULL
```

The variable "t0parsec" is not a factor so I am going to declare as a factor.

In [34]:

```
mydata6.df$t0parsec  <- factor(mydata6.df$t0parsec )
```

In [35]:

```
levels(mydata6.df$t0parsec)
```

Out[35]:

"1.1" "1.2" "2" "3" "4" "5" "6" "7"

In [36]:

```
is.factor(mydata6.df$t0parsec)
```

Out[36]:

TRUE

I now estimate a logit model of school GCSE outcomes (5+ GCSEs and grade A - C).

It will be a survey based model (svy).

Outcome variable "s15a_c".

Explanatory variables "girls", "ethnic3", "t0parsec".

In [38]:

```
mydesign5 <- svydesign(id = ~serial,data = mydata6.df, weight = ~weight)
```

In [39]:

```
model6<-svyglm (s15a_c ~ girls + factor(ethnic3) + factor(t0parsec), design
=mydesign5, data = mydata5.df, family = "binomial")
```

```
Warning message:
In eval(expr, envir, enclos): non-integer #successes in a binomial glm!
```

In [40]:

```
summary(model6)
```

Out[40]:

```
Call:
svyglm(formula = s15a_c ~ girls + factor(ethnic3) + factor(t0parsec),
    design = mydesign5, data = mydata5.df, family = "binomial")

Survey design:
svydesign(id = ~serial, data = mydata6.df, weight = ~weight)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.71584    0.09351   7.655 2.07e-14 ***
girls                0.43422    0.03999  10.858  < 2e-16 ***
factor(ethnic3)2     1.49065    0.33070   4.508 6.61e-06 ***
factor(ethnic3)3     0.59759    0.12250   4.878 1.08e-06 ***
factor(ethnic3)4     0.32020    0.30585   1.047  0.29514
factor(ethnic3)5     0.20791    0.18583   1.119  0.26323
factor(ethnic3)6     -0.71506     0.17527  -4.080 4.54e-05 ***
factor(t0parsec)1.2   0.37121     0.11991   3.096  0.00197 **
factor(t0parsec)2    -0.41717     0.10001  -4.171 3.05e-05 ***
factor(t0parsec)3    -0.84849     0.10216  -8.305  < 2e-16 ***
factor(t0parsec)4    -1.43745     0.10249 -14.025  < 2e-16 ***
factor(t0parsec)5    -1.74863     0.12005 -14.565  < 2e-16 ***
factor(t0parsec)6    -1.94399     0.10789 -18.019  < 2e-16 ***
factor(t0parsec)7    -2.36128     0.12206 -19.345  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000279)

Number of Fisher Scoring iterations: 4
```

I now pass the results to the quasi-variance procedure.

In [41]:

```
model6.qvs <- qvcalc(model6, "factor(t0parsec)")
```

In [42]:

```
summary(model6.qvs, digits = 4)
```

```
Model call:  svyglm(formula = s15a_c ~ girls + factor(ethnic3) + factor(t0p
arsec),      design = mydesign5, data = mydata5.df, family = "binomial")
Factor name:  factor(t0parsec)
       estimate     SE quasiSE quasiVar
    1.1   0.0000 0.0000 0.09167 0.008403
    1.2   0.3712 0.1199 0.07754 0.006012
    2    -0.4172 0.1000 0.04040 0.001632
    3    -0.8485 0.1022 0.04527 0.002049
    4    -1.4374 0.1025 0.04555 0.002075
    5    -1.7486 0.1201 0.07746 0.006000
    6    -1.9440 0.1079 0.05673 0.003218
    7    -2.3613 0.1221 0.08033 0.006453
Worst relative errors in SEs of simple contrasts (%):  -0.2 0.4
Worst relative errors over *all* contrasts (%):  -1.2 0.5
```

I now plot the results for "t0parsec" along with quasi-variance based 95% comparison intervals.

In [43]:

```
plot(model6.qvs)
```

Parental Social Class NS-SEC (t0parsec)

1.1 Large employers and higher managerial and administrative occupations
1.2 Higher professional occupations
2 Lower managerial, administrative and professional occupations

3 Intermediate occupations
4 Small employers and own account workers
5 Lower supervisory and technical occupations
6 Semi-routine occupations
7 Routine occupations
8 Never worked and long-term unemployed

Comments

The National Socio-economic Classification (NS-SEC) is a commonly used measure in stratification research and is the measure used in official statistics and government research in the United Kingdom. In this model I replicated the analysis of school GCSE attainment using YCS Cohort 9 through the incorporation of a parental NS-SEC measure that was derrived by Croxford et al (2007).

The analyses above have required some more data wrangling. Therefore it is prudent to save a new copy of the data.

I will take a look at the objects that are knocking around.

In [44]:

```
ls()
```

Out[44]:

"model6" "model6.qvs" "mydata5.df" "mydata6.df" "mydataew.df" "mydataew2.df" "mydesign5" "mytablenssec" "myvarsew"

To avoid confusion later and to help to keep the workflow clear I will create a new data frame "mydata6.df".

mydata6.df is a file that combines YCS cohort 9 file "ycs9sw1" [SN: 4009] and "ew_core" from Croxford (2007) [SN: 5765].

In [46]:

```
str(mydata6.df)
```

```
'data.frame':	12789 obs. of  33 variables:
 $ serial       : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight       : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex          : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c        : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58          : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth        : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe       : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls       : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1     : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white       : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ethnic2     : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3     : num  1 1 3 1 1 1 1 1 1 1 ...
 $ t0cohort    : int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 ...
 $ t0nation    : Factor w/ 3 levels "england","wales",..: 1 1 1 1 1 1 1 1 1
1 ...
 $ t0source    : chr  "ycs9" "ycs9" "ycs9" "ycs9" ...
 $ t1weight    : num  0.875 0.976 0.976 0.976 0.959 ...
 $ t1resp      : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2
2 ...
 $ t0parsec    : Factor w/ 8 levels "1.1","1.2","2",..: 1 5 5 3 5 2 6 1 4 5
...
```

In [47]:

```
save(mydata6.df,file="C:/Users/Vernon/OneDrive - University of Edinburgh/Do
cuments/ycs_9_2017/ycs9sw1_v4.rda")
```

Here I make a Stata copy of the file just in case I required it for *swivel chair* activities later in the workflow.

In [48]:

```
write.dta(mydata6.df, "C:/Users/Vernon/OneDrive - University of Edinburgh/D
ocuments/ycs_9_2017/ycs9sw1_v4.dta")
```

Producing a Data Dictionary (or Codebook)

In [1]:

```
load ("C:/Users/Vernon/OneDrive - University of
Edinburgh/Documents/ycs_9_2017/ycs9sw1_v4.rda")
ls()
ls()
```

Out[1]:

"mydata6.df"

In [2]:

```
str(mydata6.df)
```

```
'data.frame':	12789 obs. of  33 variables:
 $ serial      : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight      : num  0.875 0.976 0.976 0.976 0.959 ...
 $ sex         : Factor w/ 4 levels "not answered (9)",..: 4 3 3 3 4 3 4 3
4 4 ...
 $ s1a_c       : int  9 9 9 9 2 2 7 3 10 1 ...
 $ a58         : Factor w/ 15 levels "not answered (99)",..: 4 4 8 4 4 4 4
4 4 4 ...
 $ s1eth       : Factor w/ 9 levels "not answered (9)",..: 5 5 7 5 5 5 5 5
5 5 ...
 $ s1acqe      : Factor w/ 9 levels "not answered (9)",..: 5 5 5 5 6 6 5 6
5 6 ...
 $ pseg        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg6       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ pseg7       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls       : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1     : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ white       : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ethnic2     : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3     : num  1 1 3 1 1 1 1 1 1 1 ...
 $ t0cohort    : int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 ...
 $ t0nation    : Factor w/ 3 levels "england","wales",..: 1 1 1 1 1 1 1 1 1
1 ...
 $ t0source    : chr  "ycs9" "ycs9" "ycs9" "ycs9" ...
 $ t1weight    : num  0.875 0.976 0.976 0.976 0.959 ...
 $ t1resp      : Factor w/ 2 levels "did not respond",..: 2 2 2 2 2 2 2 2 2
2 ...
 $ t0parsec    : Factor w/ 8 levels "1.1","1.2","2",..: 1 5 5 3 5 2 6 1 4 5
...
```

In [10]:

```
myvarscb <- c("serial", "weight", "s15a_c", "girls", "ethnic1", "ethnic2",
"ethnic3",
              "white", "black", "indian", "pakistani", "bangladeshi", "chine
se", "other",
              "prof_man", "o_non_man", "skilled_man", "semi_skilled", "t0par
sec")
```

```
sec")
```

In [11]:

```
mydata7.df <- mydata6.df[myvarscb]
str(mydata7.df)
```

```
'data.frame':	12789 obs. of  19 variables:
 $ serial      : int  200001 200004 200005 200006 200014 200023 200024 2000
25 200032 200035 ...
 $ weight      : num  0.875 0.976 0.976 0.976 0.959 ...
 $ s15a_c      : num  1 1 1 1 0 0 1 0 1 0 ...
 $ girls       : num  1 0 0 0 1 0 1 0 1 1 ...
 $ ethnic1     : Factor w/ 6 levels "1","2","3","4",..: 1 1 3 1 1 1 1 1 1 1
...
 $ ethnic2     : num  4 4 3 4 4 4 4 4 4 4 ...
 $ ethnic3     : num  1 1 3 1 1 1 1 1 1 1 ...
 $ white       : num  1 1 0 1 1 1 1 1 1 1 ...
 $ black       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ indian      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ pakistani   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ bangladeshi : num  0 0 0 0 0 0 0 0 0 0 ...
 $ chinese     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ other       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ prof_man    : int  1 0 0 0 0 1 0 1 0 0 ...
 $ o_non_man   : int  0 1 0 1 0 0 0 0 1 0 ...
 $ skilled_man : int  0 0 1 0 1 0 1 0 0 1 ...
 $ semi_skilled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ t0parsec    : Factor w/ 8 levels "1.1","1.2","2",..: 1 5 5 3 5 2 6 1 4 5
...
```

Data Dictionary (or Codebook)

This is the codebook for the file **ycs9sw1_v4.rda** which contains **mydata6.df** .

---

serial id variable unique to YCS cohort 9

---

weight survey weight sweep 1 YCS cohort 9

---

s15a_c outcome variable 5+ GCSEs A (star) - C constructed from variable "s1a_c"

> 0 = 1 - 4 GCSEs grades A (star) - C
> 1 = 5+ GCSEs grades A (star) - C

---

girls gender variable constucted from variable "sex"

> 0 = boys
> 1 = girls

> 1 = girls

---

ethnic1 ethnicity variable constructed from variable "a58"

> 1 White
> 2 Black
> 3 Indian
> 4 Pakistani
> 5 Bangladeshi
> 6 Chinese
> 7 Other (but this category has been omitted from the analysis because it is omitted in Connolly 2006)

---

ethnic2 ethnicity variable constructed from variable "a58"

> 1 Black
> 2 Chinese
> 3 Indian
> 4 White
> 5 Bangladeshi
> 6 Pakistani
> 7 Others (but this category has been omitted from the analysis because it is omitted in Connolly 2006)

The variable **ethnic1** is coded to match the ethnicity measure in Table 1 (p.7) Connolly (2006). However, the order of the dummy variables included in the logistic regression model in Table 3 (p.20) Connelly (2006) do not match. The reference category for the logistic regression should be 'black' pupils (i.e. carib; afro.; other black). This could not have easily been foreseen.

---

ethnic3 ethnicity variable constructed from variable "a58"

This variable is used in the "replication" model.

The majority group 'white' pupils are now the reference category.
The three south Asian categories are adjacent to each other.

> 1 'White'
> 2 'Chinese'
> 3 'Indian'
> 4 'Banglasdeshi'
> 5 'Pakistani'
> 6 'Black'

(Others are absent from the model)

---

white dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

black dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

indian dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

pakistani dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

bangladeshi dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

chinese dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

other dummy variable constructed from variable "ethnic1"

> 0 = non-white
> 1 = white

prof_man dummy variable parents in professional / managerial social class constructed from variable "pseg1"

> 0 = no
> 1 = yes

o_non_man dummy variable parents in other non-manual social class constructed from variable

o_non_man dummy variable parents in other non-manual social class constructed from variable "pseg2"

> 0 = no
>
> 1 = yes

skilled_man dummy variable parents in skilled-manual social class constructed from variable "pseg3"

> 0 = no
>
> 1 = yes

semi_skilled dummy variable parents in semi-skilled manual social clas constructed from variable "pseg4"

> 0 = no
>
> 1 = yes

t0parsec categorical variable parents social class - a derrived variable Croxford et al. (2007) SN: 5765, The UK National Socio-economic Classification (NS-SEC) 8 category version

> 1.1 Large employers and higher managerial and administrative occupations
>
> 1.2 Higher professional occupations
>
> 2 Lower managerial, administrative and professional occupations
>
> 3 Intermediate occupations
>
> 4 Small employers and own account workers
>
> 5 Lower supervisory and technical occupations
>
> 6 Semi-routine occupations
>
> 7 Routine occupations
>
> 8 Never worked and long-term unemployed

Discussion

The Pre-Analysis Plan Reviewed

In this section I review the pre-analysis plan compare it with the work that was actually produced.

The pre-analysis plan is available here
https://github.com/vernongayle/new_rules_of_the_sociological_method/blob/master/pre_analysis_plan_

.

Tasks

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

Achieved.

2). Replication of Logistic Regression Model Reported in Connolly (2006) Using Quasi-Variance based Estimation

Achieved.

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

Achieved.

Deliverables

1). A reproducible workflow within a Jupyter notebook deposited in a Git repository
Achieved.

2). A data dictionary (codebook) accompanying the work
Achieved.

The Reproducibility Checklist Revisited

In this section I reflect on how the work compares with Stark's Reproducibility Checklist.

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

Philip Stark outlines 14 reproducibility points that an analysis can fail on

1. If you relied on Microsoft Excel for computations
Excel was not used in this work.

2. If you did not script your analysis, including data cleaning and munging
All of the analysis was scripted see Data Wrangling and Data Analysis

3. If you did not document your code so that others can read and understand it
As far as practicable I have attempted to write this Jupyter notebook as a 'literate data analysis document'.
I provided information on using this notebook, and on the authorship and meta-information.

4. If you did not record and report the versions of the software you used (including library dependencies)
I reported on the computing environment and data analysis software including library dependencies.

5. If you did not write tests for your code

5. If you did not write tests for your code
I provided two code tests, one for logistic regression and one for quasi-variance estimation, which are checked against published results.

6. If you did not check the code coverage of your tests
I did not write or use any new tests.

7. If you used proprietary software that does not have an open-source equivalent without a really good reason
The data enabling (i.e. wrangling and cleaning) and the analyses were undertaken in *R* which is an open-source software.

8. If you did not report all the analyses you tried (transformations, tests, selections of variables, models, etc.) before arriving at the one you chose to emphasize
I reported on all the analyses including data transformations, tests, selections of variables, alternative models and failed activities.

9. If you did not make your code (including tests) available
Information on how the code is licensed is provided. The code will be made available using Github https://github.com/vernongayle .

10. If you did not make your data available (and a law like FERPA or HIPPA doesn’t prevent it)
The data cannot be made publically available but researchers can assess the data from the UK Data Service https://www.ukdataservice.ac.uk/ .

11. If you did not record and report the data format
A description of the research dataset and well as information on the data format and the time and date of the dowload are provided (similar information is provided for the Croxford et al. (2007) dataset which is used to harvest an alternative social class measure.

12. If there is no open-source tool for reading data in that format
The code to read the data, wrangle the data and produce all of the results is written in *R* which is open-source and will be provided in a Jupyter notebook which is also open-source and will be made available using the open-source platform Github https://github.com/vernongayle.

13. If you did not provide an adequate data dictionary
A data dictionary (or codebook) is provided.

14. If you published in a journal with a paywall and no open-access policy
The work has not yet been published. But it will be available through UK green open access policy via my university repository http://www.research.ed.ac.uk/portal/en/persons/vernon-gayle(682d7da1-a2ad-49f0-b36c-64478c658f99).html .

Conclusions

The overall motivation of this work was to explore the practicability of using Stark’s ‘reproducibility check list’ in a piece of sociological research using genuine large-scale social science data.

The work on this project provides a striking reminder of the large amount of data enabling (i.e. data wrangling) that is required to duplicate a relatively straightforward published result. Despite knowing wrangling) that is required to duplicate a relatively straightforward published result. Despite knowing the data resource relatively well, duplicating a logit model with only three explanatory variables took me effort and some detective work. The conclusions that are drawn are the result of what is an early exploration. After further reflection and discussions they are likely to be refined. As they currently stand my conclusions are unlikely to be the last word on the subject of undertaking reproducible social science using large-scale and complex datasets.

In this section I will reflect on the items on Stark's checklist and comment on their relevance and feasibility for sociological research using large-scale social science datasets.

1. If you relied on Microsoft Excel for computations, fail.

There is little justification for using a spreadsheet to undertake analyses of large-scale social science datasets. It is almost impossible to provide and document a clear audit trail when using a spreadsheet. The now well-known case of the errors in the spreadsheet-based calculations made in Reinhart and Rogoff (2010a; 2010b) which were reported by Herndon, Ash and Pollin (2014) should serve as a stern warning against using spreadsheets in social science data analyses. In addition Stark points to the more general problems of bugs in spreadsheet software (see also http://eusprig.org/horror-stories.htm).

2. If you did not script your analysis, including data cleaning and munging, fail.

Scripting the workflow is integral to successful social science data analysis. Having a planned and organised workflow is indispensable to producing high-quality social science research. Long (2009) provides an authoritative account of good practices in the social science data analysis workflow. More recently these principles have been distilled in Gayle and Lambert (2017). In practice large-scale social science datasets are almost never delivered in an immediately 'analysis-ready' format. The data analyst will almost always have to undertake some activities to enable the data for analysis. I use the term 'data enabling' to describe the stage between downloading the social science dataset (for example from a national archive) and beginning to undertake statistical analyses. 'Data enabling' comprises tasks associated with preparing and enhancing data for statistical analysis, such as recoding measures, constructing new variables and linking datasets (Blum et al., 2009; Lambert and Gayle, 2008). 'Data enabling' is a substantial part of the research process but its importance is often overlooked. The time required to 'enable data' is frequently underestimated, even by more experienced social science data analysts. An audit trail, which acts as a set of breadcrumbs is essential for navigating back through data enabling aspects of the workflow, and is therefore essential for determining the provenance of results. A scripted workflow is essential for accurate, efficient, transparent and reproducible social science research.

3. If you did not document your code so that others can read and understand it, fail.

Documenting research code is central to delivering reproducible work. The concept of making the workflow 'literate' is new within sociological research. The idea of producing explanations of the thinking behind individual steps in the workflow is novel. Producing commentaries in human readable language (e.g. plain English) interwoven between research code and outputs is innovative. The material produced above shows promising signs that this approach will pay dividends in making research endeavours more transparent and therefore reproducible. I am mindful of the old saying 'that a recipe is not a recipe until someone else has cooked it'. A thoroughgoing proof of the literacy and the transparency of research code is only achieved when a third party, who is unconnected with the work, has successfully followed and executed the code. As a result of this position I am increasingly advocating activities such as the pair production of research code, and peer reviewing of research code. These activities will represent a marked change in how sociological research using large-scale datasets is routinely undertaken. If these activities are taken-up, and taken seriously, they will have datasets is routinely undertaken. If these activities are taken up, and taken seriously, they will have consequences for how research teams undertake work, and how researchers are trained (and re-trained).

4. If you did not record and report the versions of the software you used (including library dependencies), fail.

This is easily achieved, and can prove to be critical later when a researcher is trying to 'duplicate' the work (i.e. produce identical results). The exact results reported in table 5 Connolly (2006 p.20) could not immediately be duplicated even though identical variables were constructed. It took some detective work to ascertain that the work was undertake using SPSS in a specific mode. Since many analyses use special libraries and routines it is important that they are precisely documented so that results can be duplicated and ultimately be checked and validated.

5. If you did not write tests for your code, fail.

This is a sensible requirement, however because many sociological analysis employ standard and routine methods it may be too stringent a requirement for every single sociological analysis. In this present analysis I compared the results of two methods, which were then used in the analysis, against existing published results. Stark suggests that you should test your software every time you change it. This is a sensible and reasonable precaution, and when network versions of software are changed or updated, universities and research institutions should re-test their software.

6. If you did not check the code coverage of your tests, fail.

Stark suggests that this would be a good practice but he has never seen a scientific publication that does so. As far as I understand it, in computer science, code coverage is a measure used to describe the degree to which the source code of a program is executed when a particular test suite (a set of cases intended to be used to test a software program to show that it has some specified set of behaviours) runs. In theory a program with high code coverage has had more of its source code executed during the testing which might suggest it has a lower chance of containing undetected errors. On reflection few sociological researchers develop new statistical tests or need to implement statistical tests within new software routines. Therefore this requirement is probably irrelevant to most mainstream sociological analyses using large-scale datasets. For researchers who are developing new tests or constructing new routines then testing the coverage of their code and clearly documenting it would be a sensible action.

7. If you used proprietary software that does not have an open-source equivalent without a really good reason, fail.

It is unrealistic to undertake anything more than extremely basic analyses of survey data without using data analysis software. The requirement to use non-proprietary software however is likely to prove controversial within the community of sociological researchers using large-scale datasets.
The freeware *R* provides a viable approach with a substantial volume of analytical options and considerable programming flexibility (Long, 2011). I have shown in this analysis that *R* can be used in a standard piece of sociological inquiry. The UK Data Service currently provides datasets in SPSS and Stata format. These formats can be read in *R*. The UK Data Service provides data in a more package agnostic tab-delimited format. Some *R* users advocate importing data in this format. In my experience this format can prove challenging to work with especially when matching and merging files and undertaking data analysis enabling tasks.
I am a sociologist who has been undertaking research with large-scale and complex datasets for nearly a quarter of a century, and have taught data analysis to undergraduate and post-graduate students, early career researchers and non-academic researchers. In my experience for sociology

students the *R* learning curve is steep. The skills which are necessary to effectively exploit *R* through textual programming seem unlikely to lead to its universal adaptation amongst the wide ranging user-communities within the social sciences (see Lambert et al., 2015). A limitation is that *R* is currently not well suited to the analysis of large-scale social surveys. For example when using *R* it is difficult to effectively combine the numeric codes for variables along with both their value and variable labels. This means that users are not able to effectively exploit the meta-information on measures that is helpful for routine survey data analysis tasks. A current limitation of *R* is that there is a lack of clear and concise help files which contain applied examples that relate to the analysis of large-scale and complex social science datasets.

Within this research example I have undertaken a small amount of analysis using Python which is an emerging open-source alternative to *R*. I was unable to undertake a survey weighted analysis using a logistic regression model, but this may in part be due to my lack of competence with this software. A severe limitation of Python is that there is very little help and almost no applied examples that relate to the analysis of large-scale and complex social science datasets. At the current time there are fewer statistical routines and libraries available in Python, and Python does not offer an alternative to many packages that are available in *R*. Python is a widely used high-level programming language for general purpose programming. Python is emerging as a valuable tool in data science (e.g. for example web scrapping). In future it might unfold as a viable software for the analysis of large-scale social science datasets.

I have generally been an advocate of using Stata for the analysis of large-scale and complex social science datasets (see Gayle and Lambert 2017). Stata stands out as a sensible choice because it is a popular commercial package with a wide community of social science users. The Stata learning curve is less steep and Stata has very good documentation. Within Stata there are a wide range of analytical capabilities, and ongoing developmental activities (see Lambert et al., 2015). I have found that overall it is the single most effective and efficient tool for undertaking and successfully completing survey data analysis. The tasks associated with data enabling, exploratory data analyses, building statistical models and organising presentation-ready and publication-ready outputs (by which we mean high-quality graphs and tables of modelling results), can all be undertaken using Stata in a single uniformed environment. The development of a Stata kernel in Jupyter, and the ability to use Stata via magic cells (as demonstrated above) illustrate how the software can effectively be used within a notebook. This is attractive for developing transparent research and bundling it within a unified research object.

SPSS is a fairly ubiquitous within sociology departments. It is suited to the analysis of large-scale datasets but compared with Stata it is far more restricted in the range of statistical models that it can estimate. SPSS currently has fewer options for estimating models that are appropriate for longitudinal data. Stata is able to offer more comprehensive facilities to analyse survey datasets with complex designs and selection strategies. This is a clear benefit for social scientists working with contemporary datasets such as the UK Household Longitudinal Study (Understanding Society) and the UK Millennium Cohort Study

In practice, given the current research climate within sociology, the programing knowledge and levels of data analysis skill, the requirement to abandon proprietary software is probably too impractical a step. The requirement could be relaxed to using an established mainstream data analysis software (e.g. Stata, SPSS, *R* or SAS), but the data enabling and the data analysis must be scripted in as ‘literate’ a fashion as possible. This is essential so that a third party who is unconnected with the project can follow and understand the workflow. Where possible it would be a good practice to augment the work by reporting how an open-source analysis could be undertaken in order to assist in the duplicating (and therefore the checking) results. In practice this might mean undertaking the data enabling and analysis in Stata but documenting how the work could also be reproduced in *R* or Python.

8. If you did not report all the analyses you tried (transformations, tests, selections of variables, models, etc.) before arriving at the one you chose to emphasize, fail

models, etc.) before arriving at the one you chose to emphasize, fail.

Providing access to the complete workflow is an indispensable aspect of rendering sociological analysis transparent and reproducible. The use of Jupyter notebooks is a concrete example of organising or bundling the elements of the workflow into a 'research object' (see http://www.researchobject.org/). The use of Jupyter notebooks in sociological research extends the possibilities of material being Findable, Accessible, Interoperable and Reusable (FAIR) which is a tenet of reproducible science.

9. If you did not make your code (including tests) available, fail.

Stark states that your code should also state how it is licensed. This is a new departure in sociological research. There are a series of licenses that would be appropriate to this activity and that would chime with the wider academic ideas of attribution. In this present work I have chosen to use the MIT License. Stark further asserts that code should be published in a way that makes it easy for others to check, re-use and extend, for example by publishing it using services like Git repositories. At the current time very few sociological analyses of large-scale and complex datasets have reported all the code used to enable data and then to undertake the analysis.

Few sociological studies have used repositories. Git repositories are primarily used for source code management in software development, but can be used to keep track of changes in any set of files. These services are sometime referred to as version control software (VCS). Gentzkow and Shapiro (2014) is a rare example of VCS being recommended in the social sciences. Mercurial is an alternative to Git and, whilst GitHub has been used in this example other approaches such as BitBucket provide similar services.

10. If you did not make your data available (and a law like FERPA or HIPPA doesn't prevent it), fail.

Access to data is an integral part of transparent and reproducible social science research. The accessibility of data presents an obstacle for sociologists working with large-scale datasets. Much of the sociological analysis undertaken using large-scale and complex datasets is secondary analysis of general (or omnibus) data resources. These data resources are often national level surveys (for example the US Panel Study of Income Dynamics or the British Household Panel Survey) or data collected as part of national level Censuses. These data do not 'belong' to the data analyst and are usually provided by a national archive or other data provider under some form of 'end user license'. In practice these data are made available for research but cannot be freely shared, and all users must formally registered for the data. The rules and regulations of data use vary across countries, between data providers, and between datasets. Administrative data resources (e.g. education records) usually have tighter controls placed on their use. Sensitive or confidential data (specially relating to health) are usually especially securely controlled. Unless the data have been collected by the sociologist, and are owned and controlled by them it is unlikely that they will be able to freely share the data that have been analysed in a particular piece of work. Therefore in order to facilitate transparent and reproducible work sociologists should provide as much information on the dataset (including detailed information on versions and downloads) in order to allow a third party to get access to the data that were genuinely used in the analysis.

11. If you did not record and report the data format, fail.

In order to facilitate transparent and reproducible work sociologists should provide as much information on the dataset (including detailed information on versions and downloads) in order to allow a third party to get access to the data there were genuinely used in the analysis. This is especially important when the data are not freely available and have to be accessed via a national repository or through a data provider (see point 10 above).

12. If there is no open source tool for reading data in that format, fail.

This point is critical when datasets are being made available alongside other research objects. In short, if data are unreadable then they do not add to transparency or reproducibility. In the case of secondary analysis of existing large-scale dataset that have been provided by national data archives it is important that the code to read the data, to enable the data, and to produce all of the results is written in an accessible way. In this current project I have used *R* which is open-source and code is provided in a Jupyter notebook which is also open-source, and will be made available using the open-source platform Github https://github.com/vernongayle.

13. If you did not provide an adequate data dictionary, fail.

Providing an adequate data dictionary is a relatively easy task but it is not currently a ubiquitous practice. The acid test of a data dictionary is how easily it can be read, and how useful it is for working with the data for a third party who is unconnected with the project.

14. If you published in a journal with a paywall and no open-access policy, fail.

In the pursuit of transparent and reproducible sociological research having open access to published work is critical. Stark suggests that posting the final version of your paper on a reprint server might be enough, but he thinks that it is time to move to open scientific publications. He further states that most publishers he has worked with have let him mark up the copyright agreements to keep copyright and grant them a non-exclusive right to publish. In the context of UK higher education research, the move to Green open access will improve the accessibility of published work. Green open access involves publishing in a traditional subscription journal as usual, but also 'self-archiving' in a repository (e.g. a university archive or external subject-based repository) and providing free access (although this might be after an embargo period set by the publisher). The UK Research Council which funds research has a preference for immediate, unrestricted, on-line access to peer-reviewed and published research papers, free of any access charge and with maximum opportunities for re-use. This is commonly referred to as Gold open access (see http://www.rcuk.ac.uk/documents/documents/rcukopenaccesspolicy-pdf/).

In conclusion Stark's Reproducibility Checklist provides an important set of benchmarks, and they can reasonably be regarded as a *Berkelium Standard* (i.e. beyong gold). The items on the checklist represent solid targets to aim for. Given the present research culture in sociology, the programing skills, and the data analytical capabilities of researchers, the items on Stark's Reproducibility Checklist probably represent too large a step forward at the current time.

Therefore in the next section I posit **Some Newer Rules of the Sociological Method** which might act as a more immediate and practicable set of guidelines for undertaking reproducible sociological research using large-scale and complex social surveys and administrative datasets.

Some Newer Rules of the Sociological Method

The ultimate goal: The providence of every result should be clear and as open as possible.

The overall aim: There should be enough suitable information available to completely duplicate results, without having to contact the authors.

results, without having to contact the authors.

Here are 5 broad **'Newer Rules of the Sociological Method'** that are tailored to the analysis of large-scale and complex social science datasets.

1. Use established data analysis software (e.g. Stata, SPSS, or R), and clearly state the version, libraries, dependencies and plugins.
2. Clearly identify the version of the dataset and its origins (i.e. where and when it was obtained).
3. Write down all of the code for how the data were prepared for analysis, in a format that it can easily be read by someone unconnected with the project.
4. Write down all of the code for all of the analyses undertaken and not just the analyses that are presented, in a format that it can easily be read by someone unconnected with the project.
5. Archive the material in an accessible format at a reachable location.

> Within the archive
>
> a) Provide suitable auxillary information describing the contents of the archive, so that in future a third party unconnected with the project can understand the materials.
> b) Provide a detailed codebook.
> c) Make available all of the research code and information generated within the workflow.

The archived materials should be openly available. Try to use recognised file formats and think about how best to help a third party who is unconnected with the project understand the contents of the archive at some time in the future.

Analyzing Large-Scale and Complex Social Science Datasets

5 Simple Newer Rules of the Sociological Method

1. Tell us about your software

2. Tell us about your data

3. Show us how you got your data ready

4. Show us all the analysis you did

5. Save all of this work openly

References

Blum, J.M., Warner, G., Jones, S., Lambert, P., Dawson, A., Tan, K.L.L. and Turner, K.J., 2009. Metadata creation, transformation and discovery for social science data management: The DAMES Project infrastructure. IASSIST Quarterly, 33(1), pp.23-30.

Gayle, V. and Lambert, P., 2017. The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis, Working Paper 1/17 UK National Centre for Research Methods, http://eprints.ncrm.ac.uk/4000/.

Gentzkow, M. and Shapiro, J., 2014. Code and data for the social sciences: A practitioner’s guide, University of Chicago mimeo. Available at: https://web.stanford.edu/gentzkow/research/CodeAndData.pdf (accessed 13th December 2016).

Herndon, T., Ash, M. and Pollin, R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge Journal of Economics, 38(2), pp.257-279.

Lambert, P., Browne, W. and Michaelides, D, 2015. Contemporary developments in statistical software for social scientists, in Procter, R. and Halfpenny, P. (eds) Innovations in Digital Research Methods. London: Sage.

Lambert, P. and Gayle, V., 2008. Data management and standardisation: A methodological comment on using results from the UK Research Assessment Exercise 2008, DAMES Project Technical Paper 3.

Long, J.D., 2011. Longitudinal data analysis for the behavioral sciences using R. New York: Sage.

Long, J.S. and Long, J.S., 2009. The workflow of data analysis using Stata. College Station, TX: Stata Press.

Reinhart, C. and Rogoff, K., 2010A. Growth in a Time of Debt, Working Paper no. 15639, National Bureau of Economic Research, http://www.nber.org/papers/w15639.

Reinhart, C. and Rogoff, K., 2010B. Growth in a Time of Debt, American Economic Review, vol. 100, no. 2, 573–8

A Little Light Relief

My Jupyter Limerick

A researcher with time to fritter

Decided he didn’t need Jupyter

His results he would show

Without a traceable workflow

Could a researcher be any stupider?

Converting this Jupyter Notebook into Portable Formats

see http://nbconvert.readthedocs.io/en/latest/

1. At the cmd prompt *conda install nbconvert*
2. Change directory (for example my directory is *C:\Users\Vernon*
3. Type *jupyter nbconvert --to html mynotebook.ipynb*

The work is very exploratory.

Positive comments are always appreciated, but brickbats improve work.

or @profbigvern

END OF NOTEBOOK