

Exploring educational inequality

using census microdata

Agenda

Using census data

Census 2011 / 2021

Census microdata

Education inequality

Census 2011 / 2021

- Fieldwork in England and Wales in 2021
 - Encouraged online completion
 - Aim to deliver outputs in 2022/2023
-
- Tables
 - Geography
 - Microdata
 - Flow data

Census tables

- Standard set of univariate tables at different geographical levels
- Two / three dimensional analyses
- Promise of flexible table builder

- Incorporating controls to avoid statistical disclosure

Census geography

- Country, region and local authority
- Electoral – local authority, parliamentary constituency and ward
- Administrative geography – local authority, ward, health district ...
- Statistical geography – output areas

Census microdata

- Individual data at region and grouped local authority level (5% sample)
- Household data (1% sample)
- Secure access household data (10% sample)

Census flow data

- Origin and destination data for
 - Internal migration
 - Commuting
 - Student residence

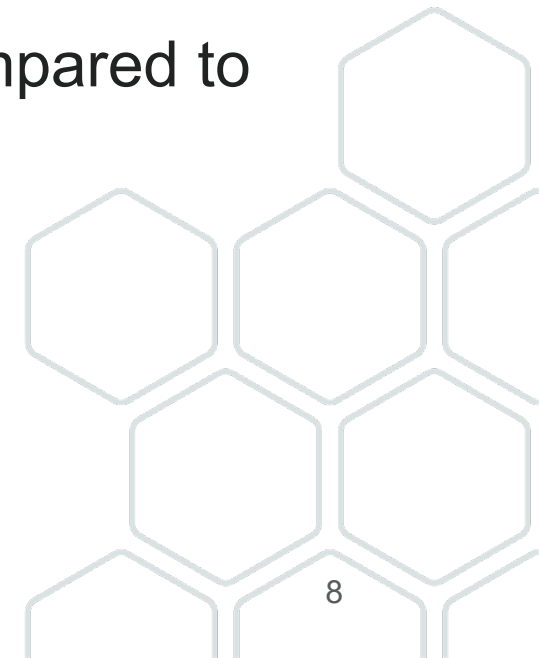
Research with Census microdata

Benefits

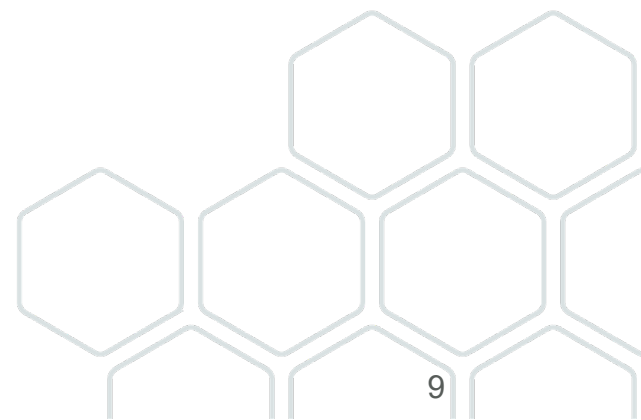
- lots of categorical variables covering demographic characteristics
- large dataset so lots of power for multivariate analysis
- comparison between local authorities / regions
- ... can add to other types of analysis using census and other data

Limitations

- no continuous / scale variables hence work needed to support linear regression
- no information on some topics e.g. crime, income
- geography limited compared to census tables



Any questions?



The data – census microdata 2011

Currently available from the UK Data Service for 1991, 2001 and 2011. The file we are using contains:

- 5% sample of census records with grouped local authority geography – 2.8 million records
- Individual and household demographic information
 - sex, age, race, migration, language, passport held, identity, religion, social class, tenure, property, household living arrangements ...
- files available with end user licence

Exploring educational inequality

With practical application using SPSS, Stata or R

Selecting a measure

Population wide measures

- Individual responses include highest level of qualification (no qualifications to degree or higher)
- Operationalised as 'having a degree'
- Excluding people under 25

- Household responses include measure of education deprivation – no-one in household having GCSE A*-C or equivalent

Expected associations

Literature suggests a number of associations with educational inequality. What aspects of an individual's characteristics do you think will be associated with having a degree?

Type of association

The following variables have been included in the model. What type of association would you expect to see?

Notes on the scripts (1)

The syntax / do / script files for SPSS, Stata and R respectively have been written to support you, both in learning how to apply the techniques required in the analysis, and, to enable you to tailor this type of analysis in the future.

The initial part of the script identifies the working directory and loads the file from a location within the file structure I used. You will need to amend these two lines in Stata and R (set working directory and get / open file) and the lines that open and save the files in SPSS to match your own.

Notes on the scripts (2)

The outputs from SPSS, Stata and R differ. For SPSS and Stata users I have used the standard output window but in R I have manipulated the output to make it readable.

I have used the chi-squared test to identify whether the association is significant ($p < 0.05$) and the strength of that association using Cramer's V as shown below:

>0.25	very strong
>0.15	strong
>0.10	moderate
>0.05	weak
>0	none

Notes on the scripts (3)

The regression outputs are with reference to the first category in all cases. To publish the information you would need to include this in the final output e.g.

Race / ethnicity

White British (reference category)

White Irish	<i>Odds ratio</i>	<i>significance</i>
-------------	-------------------	---------------------

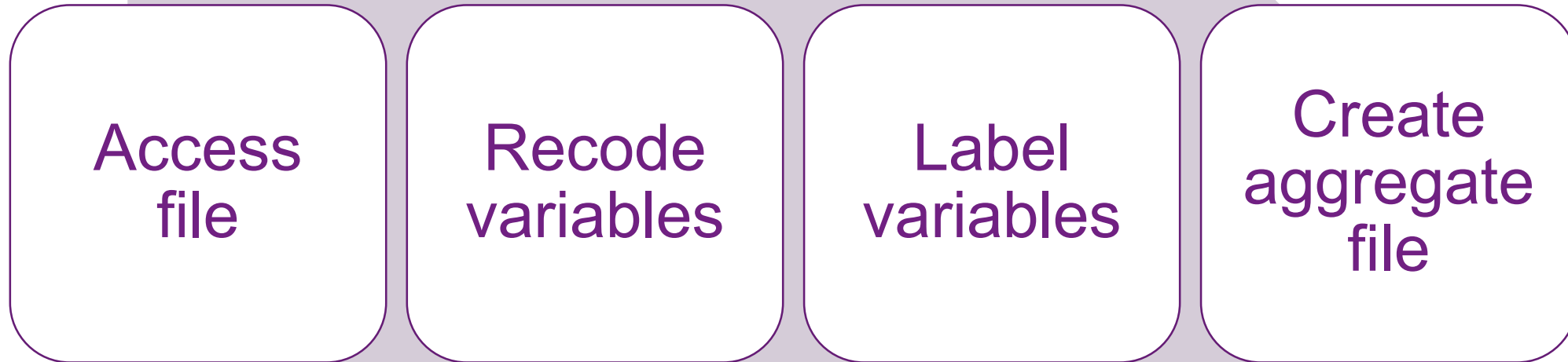
...

Notes on the scripts (4)

For the tables in R you have the option of writing them to a text file using the *sink()* command.

The regression outputs in R have been manipulated to show the coefficient label, odds ratio, lower and upper confidence intervals and the p-value. These are written to a CSV file in your working directory.

Preparing the data



Aggregate files

If you didn't manage to run the script the aggregate files are available here

<https://github.com/UKDataServiceOpen/CensusWorkshop>

Analysing educational inequality

The first stage in the script is checking the associations between having a degree and the variables selected. Based on running this part of the script, your own knowledge and examining the results, select the variables you are going to include in the model.

Running and interpreting model 1

Make any changes to the code to reflect your choices and run the first model. Identify key findings you would report.

Running and interpreting model 2

Make any changes to the code to reflect your choices and run the second model. Identify key findings you would report.



Thank you for taking part

Nigel de Noronha

nigel.denoronha@manchester.ac.uk

