Disclosure control in research environments

Felix Ritchie

Outline

- Context
- Research environments
- Use of rules
- SDC for unpredictable research
  - Principle-example teaching
  - Modelling outputs
  - Practicalities
- Example: Business Data Linking

Context

- Increasing pressure for release of microdata
  - increased ability to analyse large amounts of data
  - developments in lab technology
- Concern about release standards
  - regular SDC rules inappropriate for research
  - consistency on- and off-site
  - wider range of analyses
  - cross-border sharing
- need for clarity

Research environments

- Full access to microdata
- Expert users
- Wide and unpredictable range of outputs
- SDC carried out on outputs, not data
- Assume no deliberate attempt to remove data

Use of rules

- Consider a simple table rule

a table for release must have a frequency of at least five observations unless the data has been transformed and the resulting information does not breach confidentiality by providing information which is not available from public sources

What about identification?

SDC in the research zoo

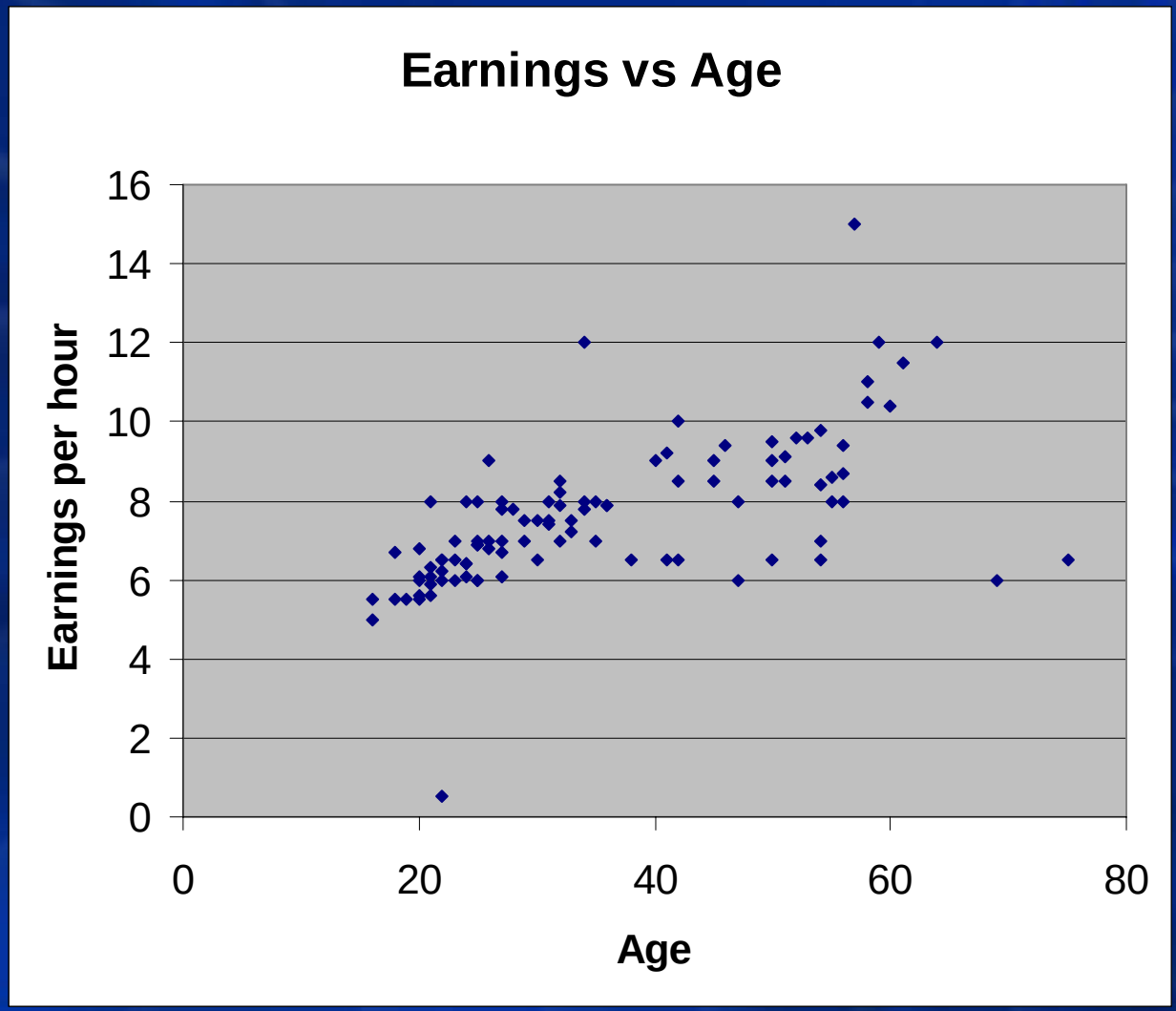
- SDC for anonymised data/aggregates developed for a finite set of outcomes
- SDC in a research environment needs to allow for an infinite set of outcomes
- Building a cage vs building a zoo

SDC for unpredictable research

- Understanding of principles
- Soft rules
  - *Table cells will normally be considered non-confidential if the frequency of units is at least five; lower frequencies can be released if it can be demonstrated that the confidentiality principles (see…) would not be broken; higher frequencies may be required if there is insufficient variation in the data or the data can be identified with a small number of statistical units*
- Modelling outputs
  - *A linear regression is non-disclosive if one or more coefficients is effectively suppressed*
- Education

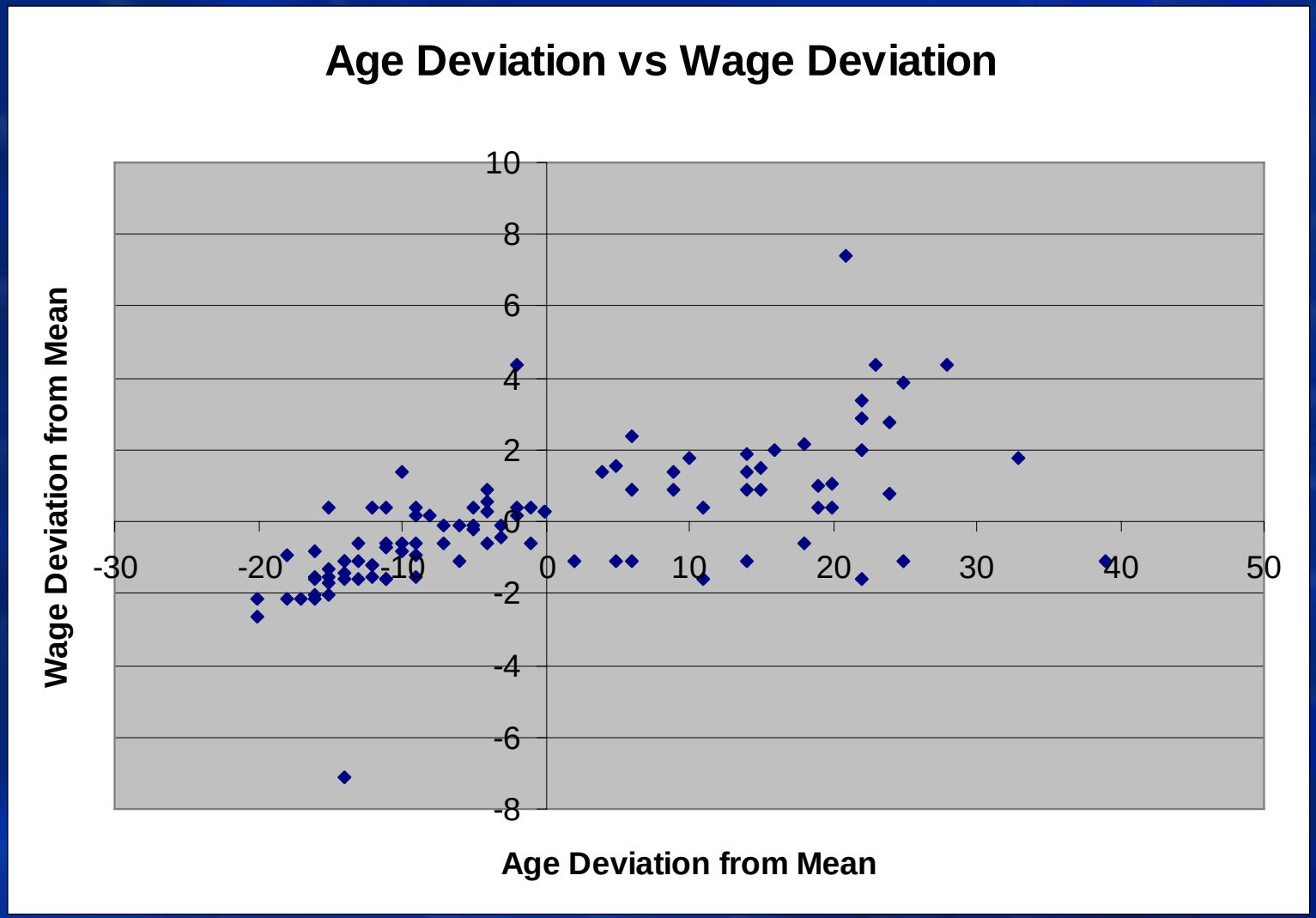
SDC teaching example

- Example: individual data



SDC teaching example

- Transformed individual data (deviation from sample mean)

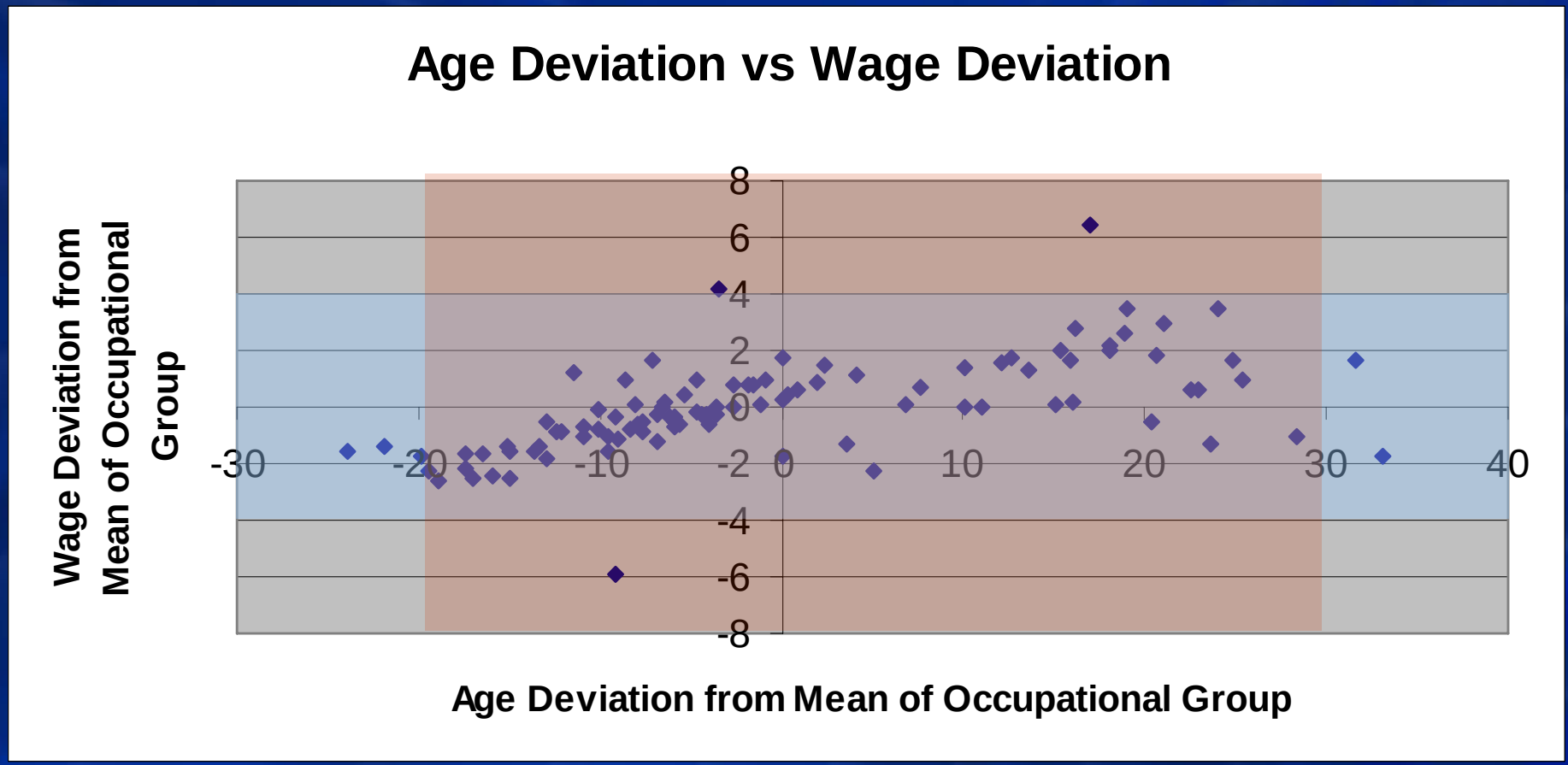


What if this graph shows nurses and hospital consultants?

What if nurses only earn £4-£8 per hour?

What if the minimum age for consultants is 40 – and they start retiring at 60?

- Transformed individual data (devi
  group mean; sub-group means ar
  observations)



Practicalities

- 4 principles put emphasis on researchers to demonstrate safety of outputs

- Potential problems
  - SDC staff motivation
  - SDC staff skills
  - Need for a positive relationship

  - Is this scalable?

Example: the UK Virtual Microdata Lab

- Compulsory/advisory training session
  - Over 300 trained researchers since 2004
- Rules developed in association with researchers
- All outputs manually reviewed
  - Target review time 5 working days
  - Practical response time 1 working day
- Quantity is main reason for rejection
  - roughly 1 output/month rejected for confidentiality reasons

Example, continued

- Problems:
  - volume of output
  - scalability
  - no SDC paper for circulation yet
- Not problems
  - skill levels
- transparency and consistency through
  - common training programme
  - recording of novel decisions

Summary

- Transparency, consistency, flexibility increasingly a concern

- Solutions:
  - Agreed principles
  - Transparency in decision-making
  - Model-based reasoning, not data-based
  - Education
  - Relationship with researchers

Felix Ritchie
felix.ritchie@ons.gov.uk

The Virtual Microdata Laboratory
vml@ons.gov.uk