Preparing wage microdata for a data explorer
SDC issues

Hans Haraldsson

Background

- Statistics Iceland is interested in making microdata on wages available to the public through a data explorer
- Would allow users to explore and visualize the distribution of wages in a give year by up to 4 background variables
  - occupation, sector, age, experience, gender, education etc.
- Some "predictors" categorical, some numerical
- Many unique combinations of values
  - Values would often be known to others
  - This would be population data so all sample uniqueness is also population uniqueness
- Clear SDC issues

Features of the SDC problem

- Explorer has no single purpose
  - Intended for "rummaging"
  - Difficult to think of a unidimensional utility measure
- Threat of identification likely the only SDC problem
  - Some groups certainly have very similar wages but these are determined by union contracts that are public
  - Other wages cannot be predicted accurately enough for inferential disclosure to be a problem
- While everything that goes on the server will ultimately be in tables the idea is to protect the microdata before the tables are prepared

Solution 1: Binning and merging

- One solution is to bin numerical variables and merge categorical variables until some criteria for k-anonymity and l-variability have been met
- A fairly "safe" solution
  - Explainable with high face validity
- Statistically problematic as distribution of values within bins/categories can differ by other bins/categories
  - Example: Female plumbers aged 20-45 are younger on average than male plumbers aged 20-45
  - Age effect on wages could appear as gender effect

Solution 1: Binning and merging cont.

- While this solution is not optimal we are making some preparations for testing it.
- We're making an implementation of the SUDA algorithm that logs all minimally unique combinations of variables for each record that is sample unique
  - Can (soon) be used with criterion >1
- We're working on a good solution to handle hierarchical codes like ISCO and ISCED
  - E.g. if gender, ISCO and ISCED make a record minimally unique it would be better to know which digit of the ISCO and ISCED classifications is needed for uniqueness
  - Allows us to test bins of various widths
- Algorithm would be used to guide binning and merging

Solution 2: Perturbation

- Additive or correlated noise
- Additive noise does not work for the general public
  - Attenuation
- We see no benefit of correlated noise over synthesis

Solution 2: Synthetic data

- From a purely statistical standpoint this solution could be ideal
- Software solutions like `synthpop` available "off the shelf"
- Finite and fairly small number of combinations of "predictors" users could use
  - It would be manageable to simply check percentiles, means, regression slopes etc. of synthetic data against real data for every combination
  - Adjust synthesis until acceptable
- Face validity would be a major issue

Solution 3: Mixed real and synthetic data

- The solution we find most interesting
- Replacing records with unique combinations of values on categorical variables with synthetic/imputed values for wages
- Smallish random sample of other records also get synthetic wage values
- Multiple synthetic values for each observation
  - Uniques with easy-to-predict wages get a small range of values, those with atypical wages get a large range
- Non synthetic observations duplicated the same number of times to approximately preserve percentiles

Solution 3: Mixed real and synthetic data cont.

- We've tried this using a tree based approach very similar to the `syn.cart()` method from `synthpop`
- Tree based methods have a number of benefits:
  - Missing data on predictors much less of a problem than with a linear model
  - Setting the penalty parameter to 0 and some minimum bucket size k basically amounts to grouping the records together in groupings of at least k records with similar values on the predictors and similar wages, then swapping values
  - First results are encouraging but there is perhaps a little too much regression to the mean in predicted values for unique records

Solution 3: Mixed real and synthetic data - numerical predictors

- Numerical variables tend to have more values than categorical ones
- So what if users want to make a scatterplot?
- We tend to think that binned scatterplots with very small bins and a minimum bin where one point represents the minimum number or less is an acceptable solution