Generating synthetic data

A example using National Pupil dataset.

Harvey Goldstein

University of Bristol

What are synthetic data?

- US Census Bureau:
- “Synthetic data are microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data is (sic) created by statistically modelling original data and then using those models to generate new data values that reproduce the original data’s statistical properties. Users are unable to identify the information of the entities that provided the original data.”
- In other words synthetic data bear no relationship to original data records except that the synthetic joint distribution closely approximates that of the original data.

The (basic) multivariate normal model

- Assume, to start, variables have a joint MVN distn.
- We use the following steps to create a basic synthetic dataset:
  - Estimate the mean and covariance matrix of the original variables. Note we can readily generalise to multilevel data and any missing values can be imputed using standard (Bayesian) procedures.
  - Formally we have $Y \sim MVN(\mu(Y_2), \Omega)$, where Y_2 is a set of 'auxiliary variables that we do not want to synthesise – i.e. they will retain their original values. In our example we assume that Y_2 is just an intercept so that we synthesise all variables and have $Y^* \sim MVN(\beta_0, \Omega)$.
  - Generate *N* data records from Y^* using estimated mean and covariance matrix. The resulting dataset satisfies criteria for being synthetic.

Analysis of synthetic data

- We fit our model of interest (MOI) to our synthetic dataset.
- Since the data have the same structure as original data (subject to the random sampling from Y^*) the parameter (point) estimates are consistent and may be used in an exploratory fashion to arrive at a final small model set that can be run with original data.
- We note that the ‘naïve’ standard errors produced will be too small. If we wish to make inferences with consistent standard errors we can generate k synthetic datasets and utilise the variation between the parameter estimates, together with the naïve standard errors, to provide the required values.

Non-normal data

- In reality we do not have MVN data, but a mixture of (normal and non-normal) continuous and categorical data.
- Goldstein et al. (2009) showed how to transform these quite generally into a MVN dataset so that the above steps can be applied and then a back transformation will supply the synthetic data on the original scales.
- This, for example, for a binary variable a ‘probit’ transformation will transform from the (0,1) scale to a standard normal distribution.

An example using a subset of NPD (2016) data

Table 1. Model fitted to original data, a single synthetic dataset and combined over 10 synthetic datasets. Standard error estimates in brackets. Number of level 1 units =9084, number of level 2 units = 464. Response is a normalised KS4 score.

Parameter	Original data	A single synthetic dataset with simple standard error estimates	Averaged over 10 datasets with augmented standard error estimates
Intercept	-4.497 (0.046)	-4.543 (0.046)	-4.504 (0.057)
KS2 score	0.955 (0.010)	0.966 (0.010)	0.957 (0.012)
female	0.179 (0.013)	0.181 (0.013)	0.188 (0.018)
English as an additional language	0.251 (0.024)	0.251 (0.023)	0.238 (0.031)
Eligible for free school meals	-0.268(0.015)	-0.271 (0.015)	-0.274 (0.019)
Level 2 variance	0.063 (0.005)	0.054 (0.005)	0.061 (0.008)
Level 1 variance	0.336 (0.005)	0.340 (0.005)	0.338 (0.006)

Conclusions, 1

- We see that the standard errors in the first two columns are extremely close.
- To use the synthetic data for inference purposes, the third column standard error estimates, which are larger, will be appropriate. For model exploration purposes, however, a single synthetic dataset will often suffice, using the sample of synthetic datasets as in column 3 as a final summary.

Conclusions, 2

- The possibility of dispensing entirely with the original data for inference purposes is an interesting one to explore.
- Clearly this would involve a loss of statistical efficiency. It would not also require access to the original data.
- It means that different data analysts supplied with different sets of synthetic samples will produce somewhat different estimates for the same MOI, which may not be desirable. To avoid this the same set of synthetic datasets could be supplied to each data analyst, for example through a publicly available depository.

Further work 1

- Practicality of procedure with large numbers of variables.
- Practicality with very large datasets – subsampling is possible
- In what circumstances can analysts legitimately make (less efficient) valid inferences without recourse to original data.
- Provision of efficient software. Currently STATJR, JOMO and MLwiN are used.
- If data confidentiality is a persistent issue it would also be possible to further anonymise synthetic data e.g by adding noise. This, however, is strictly not necessary.
- Computational problems may arise when a very large number of variables are to be synthesised.

Further work 2

- Interactions remain a problem:
    - We can form a series of (unordered) categorical variables by expanding each interaction term, with a categorisation of any continuous variables involved.
    - These can be incorporated in the synthesising.
    - Alternatively the synthetic data can be produced by running the synthesis model jointly with a MOI that contains a superset of interaction terms that will be used by the analyst. (model compatible synthesis).
    - This requires a dialogue between analysts and data providers and could be semi-automated with requests for further synthesised data simply involving the specification of the required interaction (or polynomial) terms.