Internet use in the Scottish Household Survey (2001/2002).

Details of the variables are

This data set consists of records from the Scottish Household Survey looking at factors determining internet use by Scottish adults .Permission to make this data set available were obtained as part of the PEAS project on survey research methods, see here. It uses interviews carried out in **2001/2002** with data from the Random Adult data set from this survey. There were 28 685 respondents in these two years.

The data is an R data frame ***shs*** with the following variables:saved as shs.Rdata

All categorical variables have been made into factors with levels assigned.

Output from codebook.syn(shs)

	variable	class	nmiss	perctmiss	ndistinct	details
1	shs_6cla	factor	46	0.16	6	Urban Rural classification
2	council	factor	0	0	32	See table in labs
3	hours_int	factor	19823	69.11	6	See table in labs
4	int_grocery	factor	19823	69.11	3	Use internet for groceries 'no' 'yes' 'no internet'
5	int_other	factor	19823	69.11	3	Use internet for other purchases 'no' 'yes' 'no 'internet'
6	intuse	factor	0	0	2	'no' 'yes'
7	groupinc	factor	0	0	6	Household income (grouped)
8	age	numeric	1	0	75	Range: 16 - 90
9	sex	factor	0	0	2	'male' 'female'
10	emp_sta	factor	0	0	12	Employment status See table in labs

Practical 2 b)

The practical consists of using different methods to synthesise this data set. You can do it any way you want but here are some suggestions, code following them is in **synthesise_shs.R.**

Suggested tasks

- Take a look at the data and try to understand the variables a little. How did 2001/2 internet use compare to now? Use codebook.syn to check the data.
- Are there any variables with many levels that you might want to put at the end of the visit.sequence or drop all together?
- Look at the tables of variable "intuse" by other internet variables. Are there rules that the synthetic data should obey.
- Choose a full conditional synthesis method (e.g. cart, ctree , parametric) this method. Perhaps change parameters to speed up synthesis and/or change utility.
- Use the functions *compare.synds* and *utility.tables* to evaluate the synthesis. Are rules preserved with the synthetic data?
- Synthesise your data with the method catall.
- Modify your synthesis with catall by increasing catall.nprior until you get some some rules that are broken
- Now use the parameter catall.structzero to force the rules to be obeyed.
- Now modify catall to make your synthesis DP with the parameter catall.epsilon (try different values)
- To make it DP you need to make a new data set where "age" is stored as an age group with function numtocat.syn(). Then synthesise this.
- What happens to utility and rules