# Third homework assignment

In this exercise you will generate different synthetic versions of the cps data and a subset of the National Health Interview Survey (NHIS) using the package synthpop in R.

1.) Read the synthpop documentation and reference manual which are available as part of the material for this homework assignment.

The file nhis.RData contains a subset of the variables and records of the NHIS data included in the R package PracTools (the dataset is called nhis.large in this package). The dataset is from the wave 2003 of the survey and contains the following variables:

| | |
|---|---|
| sex | Gender (1 = male; 2 = female) |
| age.grp | Age group (1 = < 18 years; 2 = 18-24 years; 3 = 25-44 years; 4 = 45-64 years; 5 = 65+) |
| educ | Highest level of education attained (1 = High school graduate, graduate equivalence degree, or less; 2 = Some college; 3 = Bachelor’s or associate’s degree; 4 = Master’s degree or higher) |
| race | Race (1 = White; 2 = Black; 3 = All other race groups) |
| inc.grp | Family income group (1 = < $20K; 2 = $20000-$24999; 3 = $25000-$34999; 4 = $35000-$44999; 5 = $45000-$54999; 6 = $55000-$64999; 7 = $65000-$74999; 8 = $75K+) |
| medicaid | Covered by medicaid, a governmental subsidy program for the poor (1 = Yes; 2 = No) |
| notcov | Not covered by any type of health insurance (1 = Yes; 2 = No) |

2.) Using synthpop, generate synthetic data for the variables age.grp, race, inc.grp, and notcov using CART models without smoothing. Make sure that the order in which the variables are synthesized is race, then age.grp, then inc.grp, then notcov. Also ensure that all variables which are not synthesized, are used as predictors in all synthesis models. Generate m=10 synthetic datasets and call this synthetic data object syn.data1.

3.) Synthesize the same four variables using parametric models for each variable. Generate m=10 synthetic datasets and call this synthetic data object syn.data2.

4.) Use the compare command provided in synthpop to evaluate the quality of your synthetic data for the four variables for both datasets. Would you prefer one of the two synthesis strategies based on your findings?

5.) Crosstabulate the variables medicaid and notcov using the original data and your synthetic datasets. For simplicity, you can use only the first replicate (i.e., the first synthetic dataset from syn.data1 and syn.data2) for the crosstabulation (Note: the generated datasets are stored as a list in syn.data1$syn. Thus, syn.data1$syn[[1]] will contain your first generated synthetic dataset, syn.data1$syn[[2]] will be your second dataset etc.)

6.) Use the “rules” and “rvalues” option in the syn command to ensure that the problems you detected in Exercise 5 no longer occur. Generate m=20 for both the CART and the parametric approach. Call the resulting objects syn.data3 and syn.data4, respectively. Run the crosstabulations from Exercise 5 again on both datasets to ensure that the problem no longer occurs.

7.) Run a logit model regressing notcov on sex, educ, race, and inc.grp using syn.data3 and syn.data4.

8.) Compare the results of these regression models to the results that you would have obtained based on the original data using the compare command. (Hint: make sure that synthpop uses the combining rules for partially synthetic data and that the variance estimates are appropriate for making inferences regarding the true values in the population. If you are unsure how to do this, read the vignette synthpop_inference, which is included in the material for this homework assignment.

9.) A simple risk measure for attribute disclosure risk computes, how often the reported value in the protected data is equal to the reported value in the original data. Compute this risk measure for the variable inc.grp for syn.data3 and syn.data4. Since you have multiple synthetic datasets, compute the risk measure as the average across the 20 replicates. Here is what you need to do:

   1. Note that the records in the synthetic datasets stored in syn[[1]],...,syn[[20]] are still in the same order as in the original data.
   2. Thus, for each synthetic dataset you only need to compute, how often the value for inc.grp is equal to the value reported in the original data.
   3. Once you computed this number for all 20 datasets, simply take the average.

   Do you think the risk of attribute disclosure is large? What would be the expected number of unchanged values, if you would generate synthetic income group values by simply assigning one of the income classes totally at random?

In the following exercises, you should use the cps data again (i.e., you need to load cps5000.RData).

10.) Synthesize only the tax variable using a CART model with a kernel density estimator applied when drawing synthetic values from the leaves of the regression tree. Again, make sure that you use all other variables as predictors when building the tree. Generate m=10 synthetic datasets and call the synthetic data object syn.data5

11.) Synthesize the tax variable using fully parametric methods (note: normrank is not fully parametric). Transform the continuous variables if necessary, but ignore for now that you might have spikes at zero or that your synthetic data might contain negative values. Generate m=10 synthetic datasets and call the synthetic data object syn.data6.

12.) Compare the tax variable in your synthetic datasets with the tax variable in the original data using the compare command. Which problems do you notice?

13.) Generate a synthetic dataset based on parametric models that accounts for the large number of zeros in the tax variable (hint: look for semicontinuous in the reference manual). Call this dataset syn.data7. Check whether syn.data7 captures the spike at zero present in the original data.