Societal Impacts of Data
Statistics 184
Professor: Mark S. Handcock

Homework 3

Due date on the Bruinlearn *Assignments* page

1) This question is about individual risk assessment of disclosure from a released data set. Take the `eusilc` data set from the `R` package `laeken`. Assume the following disclosure scenario that defines age, `pb220` (citizenship), `pl030` (education level), `rb090` (gender) and `hsize` (household size) as categorical key variables. We will use the `R` package `sdcMicro` to do the analysis. There is detailed information on the CRAN website.

a) Use the `R` package `sdcMicro` to create an object of class `sdcMicroObj` including the sampling weights (function argument `weightVar` in `createSdcObj`) and the household ID (function argument `hhId`). For information on the household ID and individual sampling weights variables, look at the manual page for the data set `eusilc`.

b) Compute the individual risk using a command like:

```
risk <- get.sdcMicroObj(sdc, type="risk")$individual
```

where `sdc` is the `sdcMicroObj`. Plot the distribution of the individual risk. Describe the results?

c) Extract the household risk and plot the distribution of the household risk. What do you see? Are some estimated risks too high?

d) Compare the individual risks to the household risks? Are the household risks higher than the individual risks?

e) Estimate the global risk for the data set above. Then, use a 10% subset of the data set and compare the results on the global disclosure risk. Do smaller data sets imply higher risks?